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° 3

OBJECTIF DE L'ACTIVITE PRATIQUE :

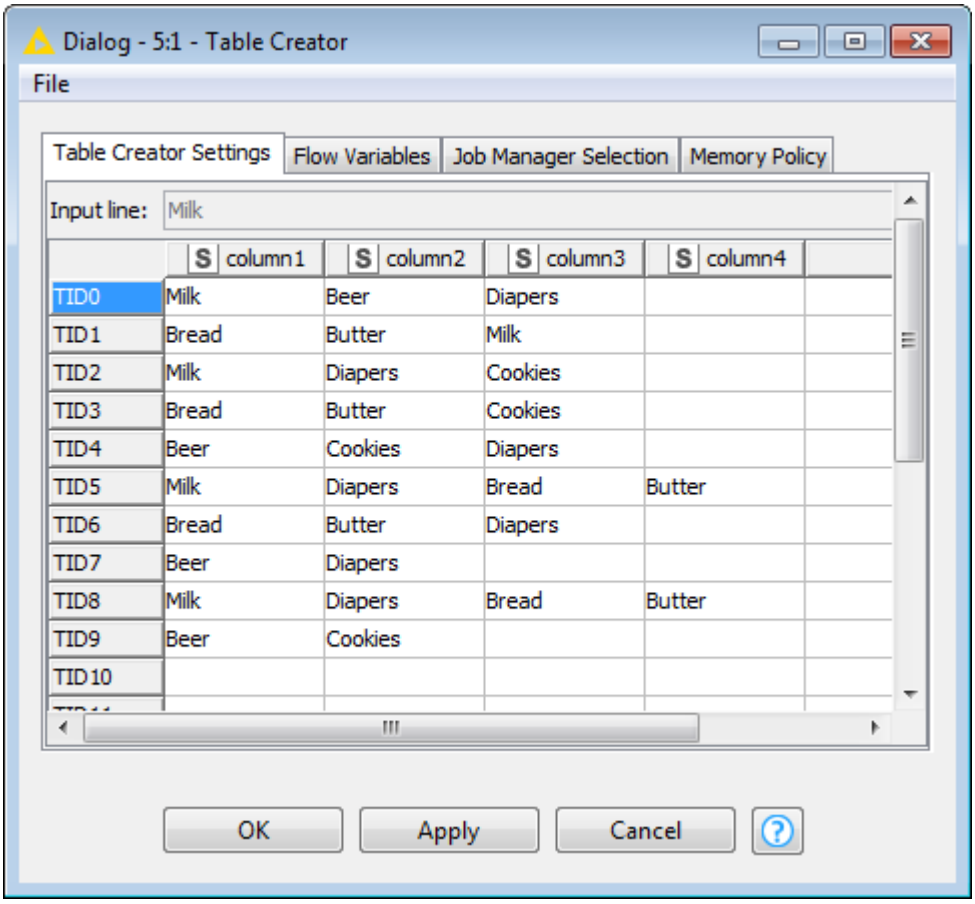
*Dans ce TP vous allez manipuler le logiciel **KNIME** pour calculer des règles d'associations.*

Création de la base de données transactionnelle

Considérons la BDT suivante :

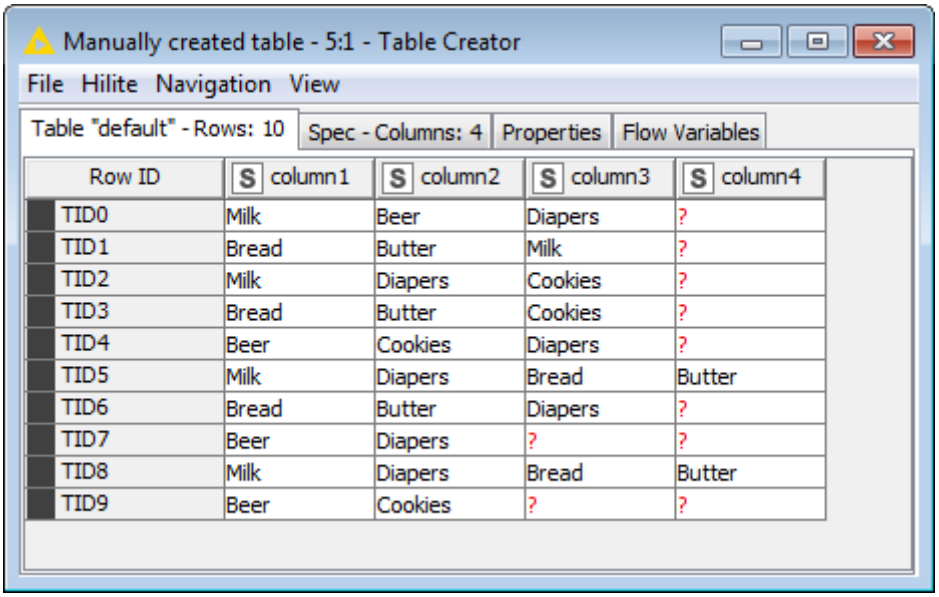
Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

1. Lancez **KNIME** (assurez-vous qu'il démarre avec un workflow vide).
2. Ajoutez un nœud « **Table Creator** » au workflow (**IO** / **Other** / **Table Creator**). Ensuite, saisissez les données relatives à chaque transaction (une transaction par ligne) comme le montre la figure 1.
3. Exécutez le nœud, puis visualisez la table construite. Remarquez la présence de **valeurs manquantes** (figure 2). Ce n'est pas grave dans ce cas. Elles seront traitées pendant la configuration du prochain nœud du workflow.



	column1	column2	column3	column4
TID0	Milk	Beer	Diapers	
TID1	Bread	Butter	Milk	
TID2	Milk	Diapers	Cookies	
TID3	Bread	Butter	Cookies	
TID4	Beer	Cookies	Diapers	
TID5	Milk	Diapers	Bread	Butter
TID6	Bread	Butter	Diapers	
TID7	Beer	Diapers		
TID8	Milk	Diapers	Bread	Butter
TID9	Beer	Cookies		
TID10				

Fig. 1. Contenu de la BDT.



Row ID	column1	column2	column3	column4
TID0	Milk	Beer	Diapers	?
TID1	Bread	Butter	Milk	?
TID2	Milk	Diapers	Cookies	?
TID3	Bread	Butter	Cookies	?
TID4	Beer	Cookies	Diapers	?
TID5	Milk	Diapers	Bread	Butter
TID6	Bread	Butter	Diapers	?
TID7	Beer	Diapers	?	?
TID8	Milk	Diapers	Bread	Butter
TID9	Beer	Cookies	?	?

Fig. 2. Visualisation de la table de la BDT.

Création d'une colonne de type « Collection » par agrégation des colonnes

Le nœud « **Create Collection Column** » regroupe plusieurs colonnes dans une nouvelle colonne distincte de type « **collection** ». Les cellules de la nouvelle colonne sont des collections de cellules typées. Cela veut dire que

le contenu peut être divisé en toute sécurité dans le contenu de la colonne d'origine. L'opération inverse est disponible dans le nœud « **Split Collection Column** ».

4. Ajoutez un nœud « **Create Collection Column** » au workflow (**Manipulation** / **Column** / **Splite & Combine** / **Create Collection Column**).
5. Reliez le nœud « **Create Collection Column** » à la sortie du nœud précédent :

Fig. 3. Les deux nœuds reliés.

Voici la forme de la boîte de configuration de ce nœud :

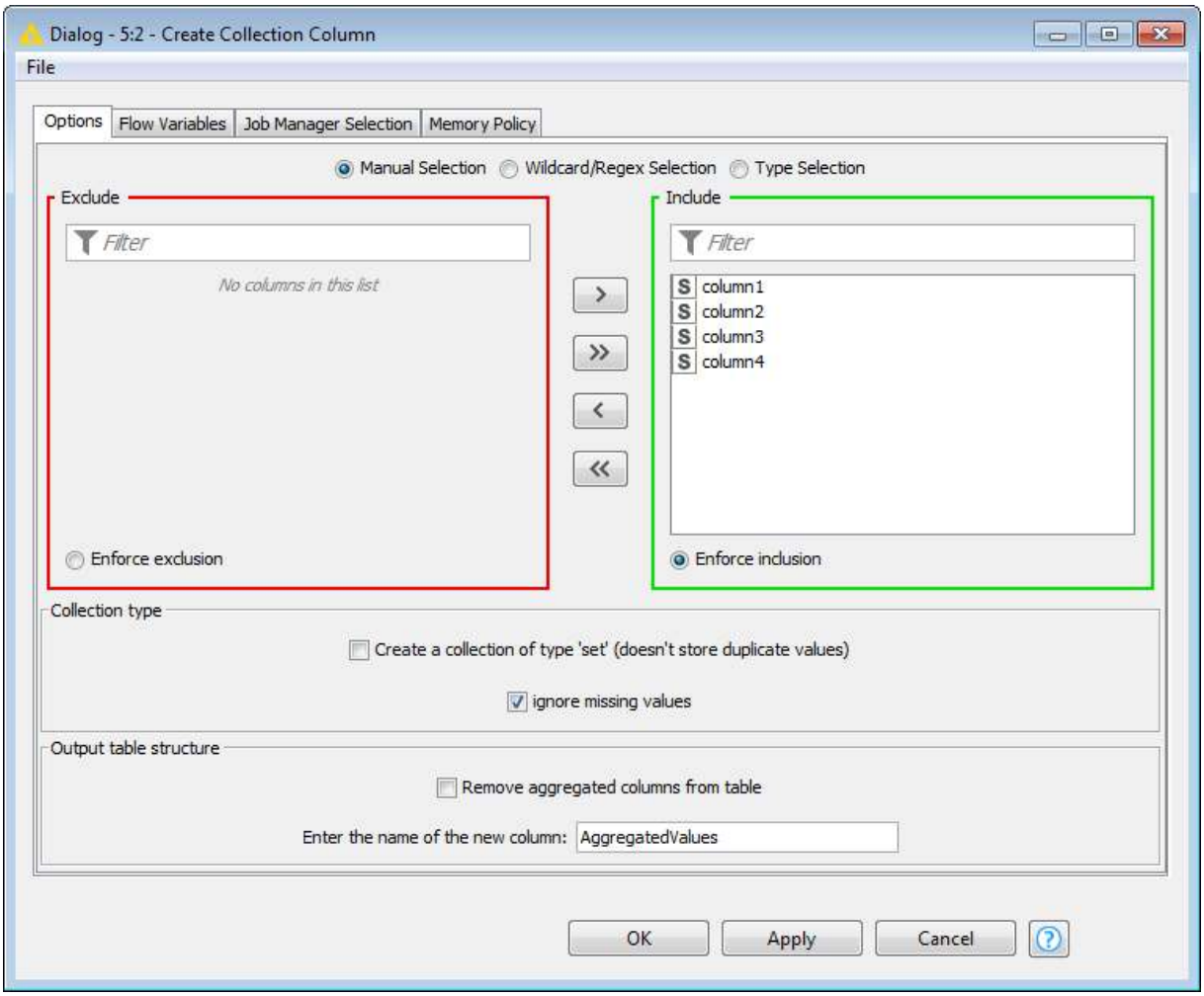
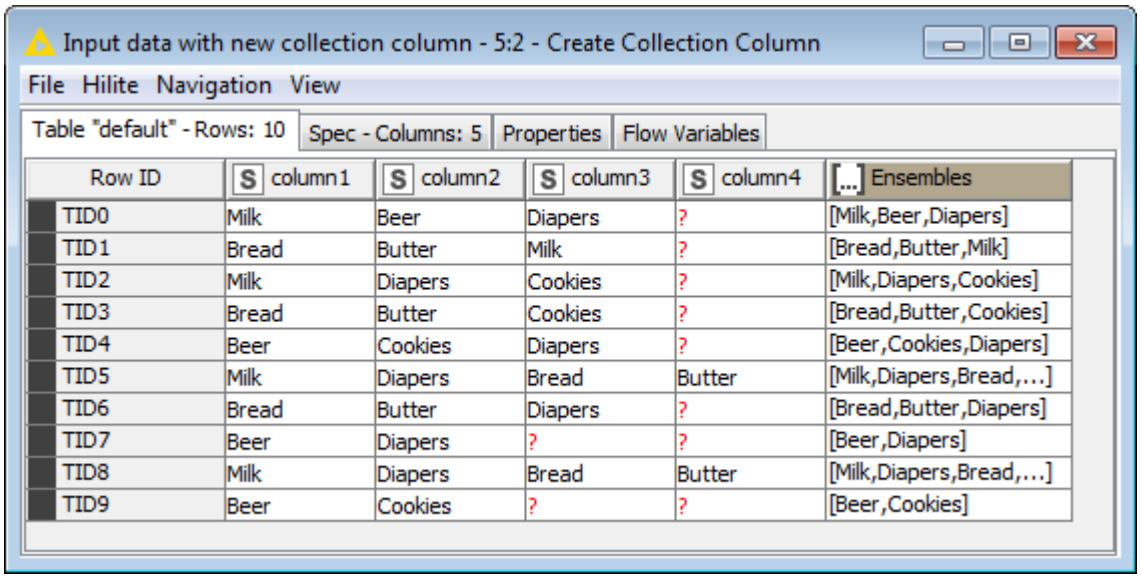


Fig. 4. Ecran de configuration du nœud « Create Collection Column ».

6. Configurez le nœud « **Create Collection Column** » en précisant les éléments suivants :
    i. Sélectionnez les noms de colonnes qui sont regroupés
    ii. Sélectionnez les noms de colonnes à exclure du regroupement.
    iii. Précisez le type de la collection : cochez l'option « **créer une collection de type 'set'** » (ne stocke pas les valeurs en double).
    iv. Ignorez les valeurs manquantes. Dans ce cas, les valeurs manquantes seront non stockées dans les cellules de collecte.
    v. Précisez la structure de la table de sortie : (1) Si la case « supprimer la colonne agrégée » est cochée, les colonnes agrégées seront supprimées de la table de sortie, (2) Spécifiez le nom (par exemple, "Ensembles") de la nouvelle colonne contenant les cellules de collection.

Vous obtiendrez alors le résultat suivant après exécution du nœud :



Row ID	column1	column2	column3	column4	Ensembles
TID0	Milk	Beer	Diapers	?	[Milk,Beer,Diapers]
TID1	Bread	Butter	Milk	?	[Bread,Butter,Milk]
TID2	Milk	Diapers	Cookies	?	[Milk,Diapers,Cookies]
TID3	Bread	Butter	Cookies	?	[Bread,Butter,Cookies]
TID4	Beer	Cookies	Diapers	?	[Beer,Cookies,Diapers]
TID5	Milk	Diapers	Bread	Butter	[Milk,Diapers,Bread,...]
TID6	Bread	Butter	Diapers	?	[Bread,Butter,Diapers]
TID7	Beer	Diapers	?	?	[Beer,Diapers]
TID8	Milk	Diapers	Bread	Butter	[Milk,Diapers,Bread,...]
TID9	Beer	Cookies	?	?	[Beer,Cookies]

Fig. 5. La colonne de type Collection est crée.

Analyse des règles d'associations

7. Ajoutez un nœud « **Association Rule Learner** » au workflow (**Analytics** / **Mining** / **Item Sets & Association Rules** / **Association Rule Learner**), puis le relies à la sortie du noeud « **Create Collection Column** » comme indiqué dans la figure suivante :

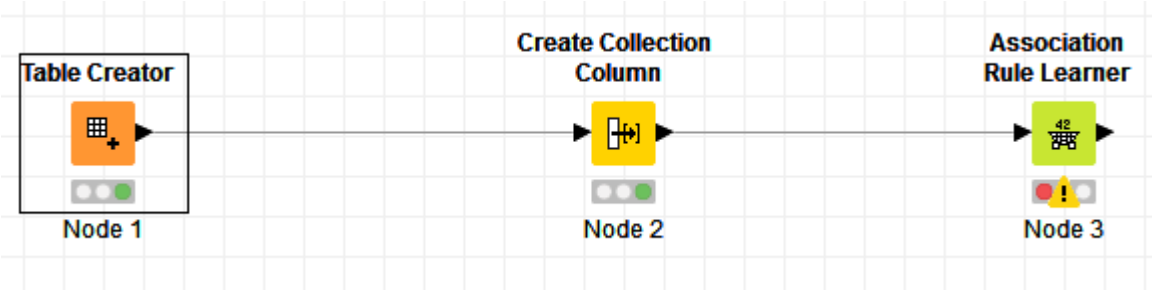


Fig. 6. Les trois nœuds du projet.

8. Configurez le nœud « **Association Rule Learner** », en utilisant la boîte suivante :

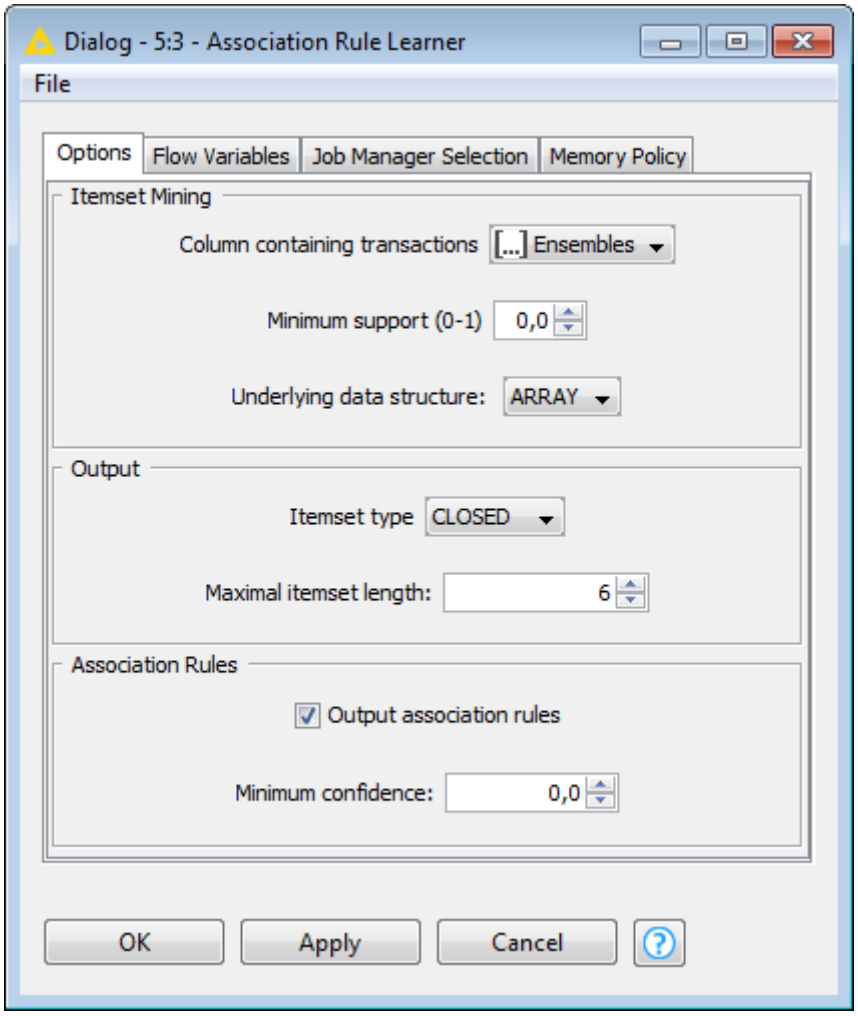


Fig. 7. La boîte de dialogue du nœud « Association Rule Learner ».

Voici comment configurer ce nœud :

i. Spécifiez la colonne contenant les transactions à explorer pour les itemsets fréquents ou les règles d'association.

ii. Précisez le support minimum (un réel entre 0 et1).

iii. Précisez la structure de données sous-jacente : **ARRAY** ou **TIDList**. **ARRAY** est recommandé lorsque le nombre de transactions (lignes) est supérieur au nombre d'éléments, et **TIDList** si le nombre de

lignes est petit et le nombre d'éléments grand. En général, l'option **ARRAY** nécessite plus de mémoire mais elle est plus rapide, tandis que la liste **TIDList** nécessite moins de mémoire mais elle est plus lente.

iv. Spécifiez le type d'itemset : **libre** (produit des itemsets redondants) **fermé** (fournissent le plus d'informations) ou **maximal** (peut masquer certaines informations).

v. Spécifiez la taille maximale des itemsets.

vi. Indiquez si l'affichage des règles d'association aura lieu (ou non). Les règles d'association sont toujours générées à partir d'items fréquents libres et sont **obligées de n'avoir qu'un seul élément dans la conséquence**.

vii. Précisez la confiance minimale.

9. On suppose que le seuil du support est fixé à 30% et que le seuil de confiance est de 40%. Répondre aux questions suivantes :

   a. Quel est le nombre d'itemsets fréquents ?

   b. Donner le support minimum et le support maximum.

   c. Déterminer trois items x, y et z tels que les deux règles $\{x\} \rightarrow \{y\}$ et $\{y\} \rightarrow \{z\}$ soient de confiance $\geq 40\%$, mais la règle $\{x\} \rightarrow \{z\}$

Voici ce que vous devez obtenir comme résultat pour cette question :

Frequent itemsets/Association rules - 5:3 - Association Rule Learner

File Hilite Navigation View

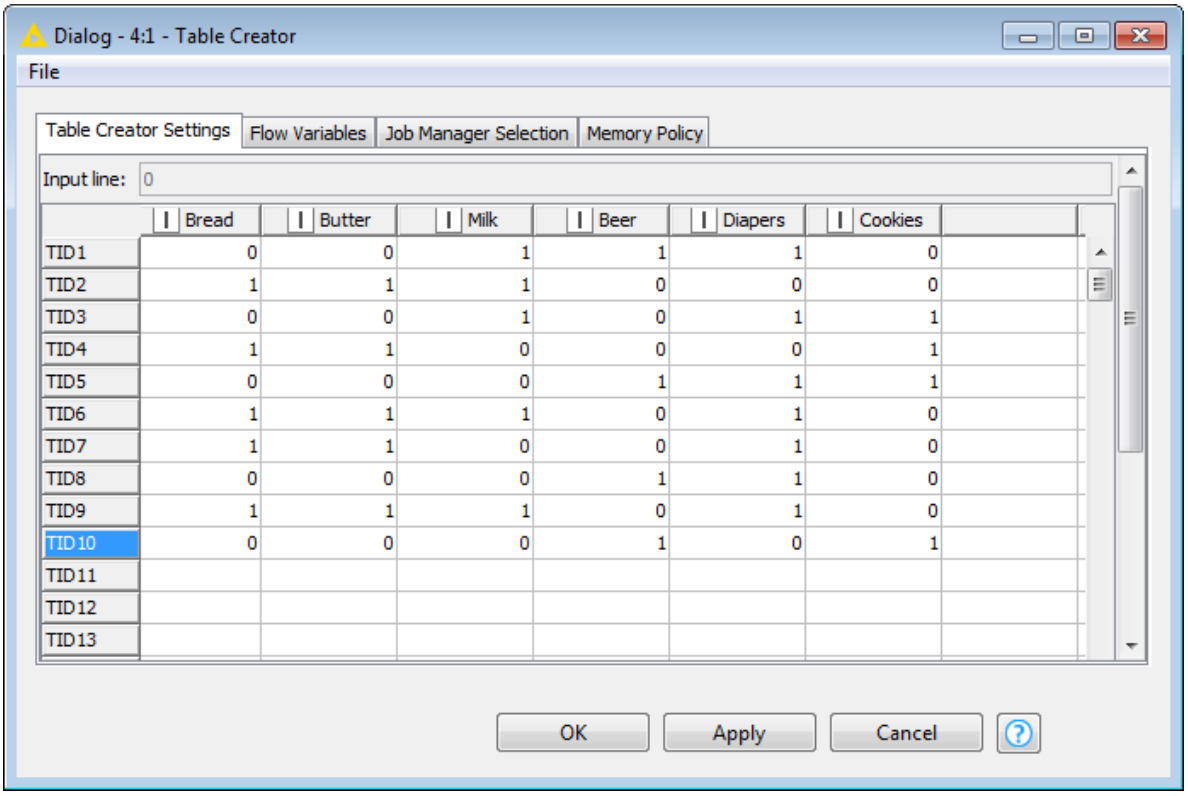
Table "default" - Rows: 12 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.3	0.429	1.071	Beer	<---	[Diapers]
rule1	0.3	0.75	1.071	Diapers	<---	[Beer]
rule2	0.3	0.6	1.2	Milk	<---	[Butter,Bread]
rule3	0.3	1	2	Bread	<---	[Butter,Milk]
rule4	0.3	1	2	Butter	<---	[Milk,Bread]
rule5	0.3	0.6	0.857	Diapers	<---	[Butter,Bread]
rule6	0.3	1	2	Bread	<---	[Butter,Diapers]
rule7	0.3	1	2	Butter	<---	[Diapers,Bread]
rule8	0.4	0.571	1.143	Milk	<---	[Diapers]
rule9	0.4	0.8	1.143	Diapers	<---	[Milk]
rule10	0.5	1	2	Bread	<---	[Butter]
rule11	0.5	1	2	Butter	<---	[Bread]

Fig. 8. Résultat obtenu pour la question 9.

Création de la base de données transactionnelle sous la forme binaire

1. Créez un nouveau workflow.
2. Ajoutez un nœud « **Table Creator** » au workflow (**IO** / **Other** / **Table Creator**), puis le-configurez de façon à obtenir le résultat suivant :



	Bread	Butter	Milk	Beer	Diapers	Cookies
TID1	0	0	1	1	1	0
TID2	1	1	1	0	0	0
TID3	0	0	1	0	1	1
TID4	1	1	0	0	0	1
TID5	0	0	0	1	1	1
TID6	1	1	1	0	1	0
TID7	1	1	0	0	1	0
TID8	0	0	0	1	1	0
TID9	1	1	1	0	1	0
TID10	0	0	0	1	0	1
TID11						
TID12						
TID13						

Fig. 9. Création d'une BDT sous forme formelle (Chaque item est représenté par une variable binaire).

3. Ajoutez un nœud « **Create Bit Vector** » au workflow (**Analytics** / **Mining** / **Item Sets & Association Rules** / **Create Bit Vectior**). Reliez ce nœud à la sortie du nœud précédent.
4. Ajoutez un troisième nœud « **Association Rule Learner** » et le-reliez à la sortie du second nœud déjà crée. Vous devez avoir le schéma suivant :

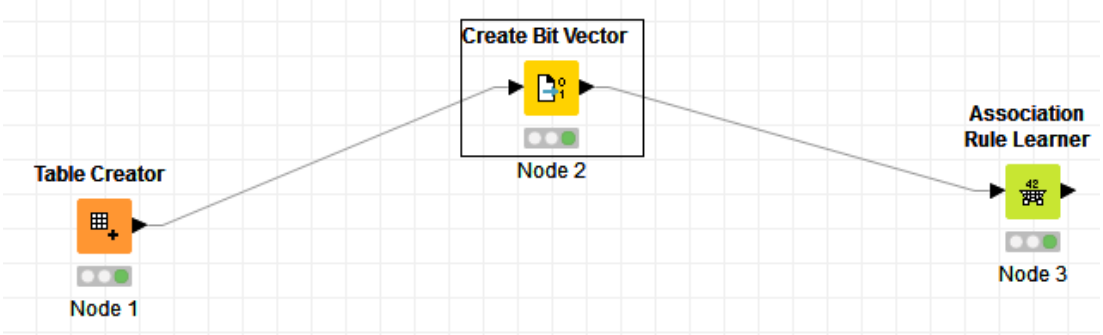


Fig. 10. Les nœuds du projet.

5. Configurez le nœud « **Create Bit Vector** » en cochant l'option « **Create bit vector from multiple numeric columns** » :

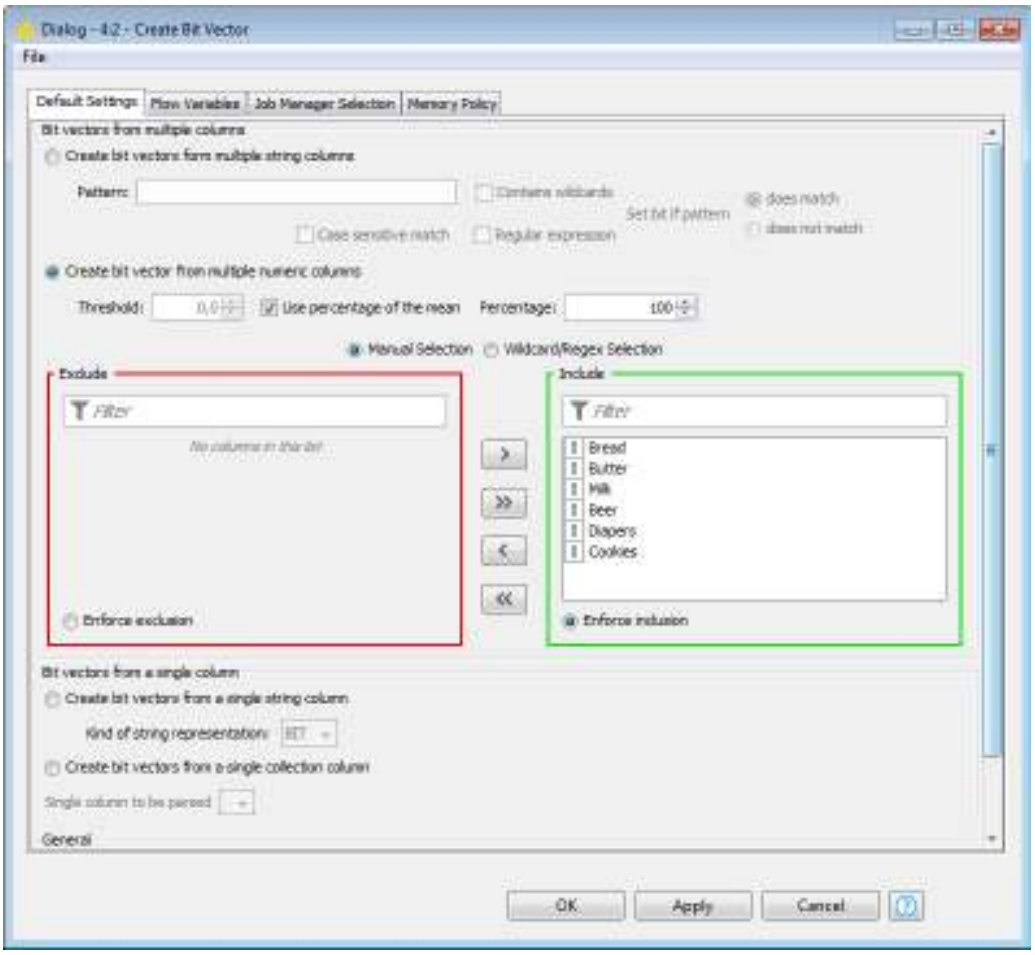


Fig. 11. Configuration d'un nœud « Create Bit Vector » .

6. Configurez le nœud « Association Rule Learner » pour répondre à la question 9 du premier projet. Comparez les résultats obtenus.

Frequent itemsets/Association rules - 4:3 - Association Rule Learner

File Hilite Navigation View

Table "default" - Rows: 12 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Support	Confide...	Lift	Conseq...	implies	Items
rule0	0.3	0.429	1.071	Beer	<---	[Diapers]
rule1	0.3	0.75	1.071	Diapers	<---	[Beer]
rule2	0.3	1	2	Bread	<---	[Butter,Milk]
rule3	0.3	1	2	Butter	<---	[Bread,Milk]
rule4	0.3	0.6	1.2	Milk	<---	[Butter,Bread]
rule5	0.3	1	2	Bread	<---	[Butter,Diapers]
rule6	0.3	1	2	Butter	<---	[Diapers,Bread]
rule7	0.3	0.6	0.857	Diapers	<---	[Butter,Bread]
rule8	0.4	0.571	1.143	Milk	<---	[Diapers]
rule9	0.4	0.8	1.143	Diapers	<---	[Milk]
rule10	0.5	1	2	Bread	<---	[Butter]
rule11	0.5	1	2	Butter	<---	[Bread]

Fig. 12. Résultats obtenus en utilisant une BDT formelle.