Exo1 : Prise en main de RapidMiner pour l'extraction de règles d'association

RapidMiner est une suite logicielle permettant d'effectuer plusieurs tâches d'extraction de connaissance, en particulier, la recherche d'association, la classification et le regroupement qui sont les 3 thème qui seront abordés dans le cours.

Pour les règles d'association, RapidMiner implémente l'algorithme FP-Growth (basé sur les FP-trees vus en cours) qui permet de trouver les ensembles fréquents. Une fois les itemsets fréquents sont extraits, les règles d'association peuvent alors en être générées.

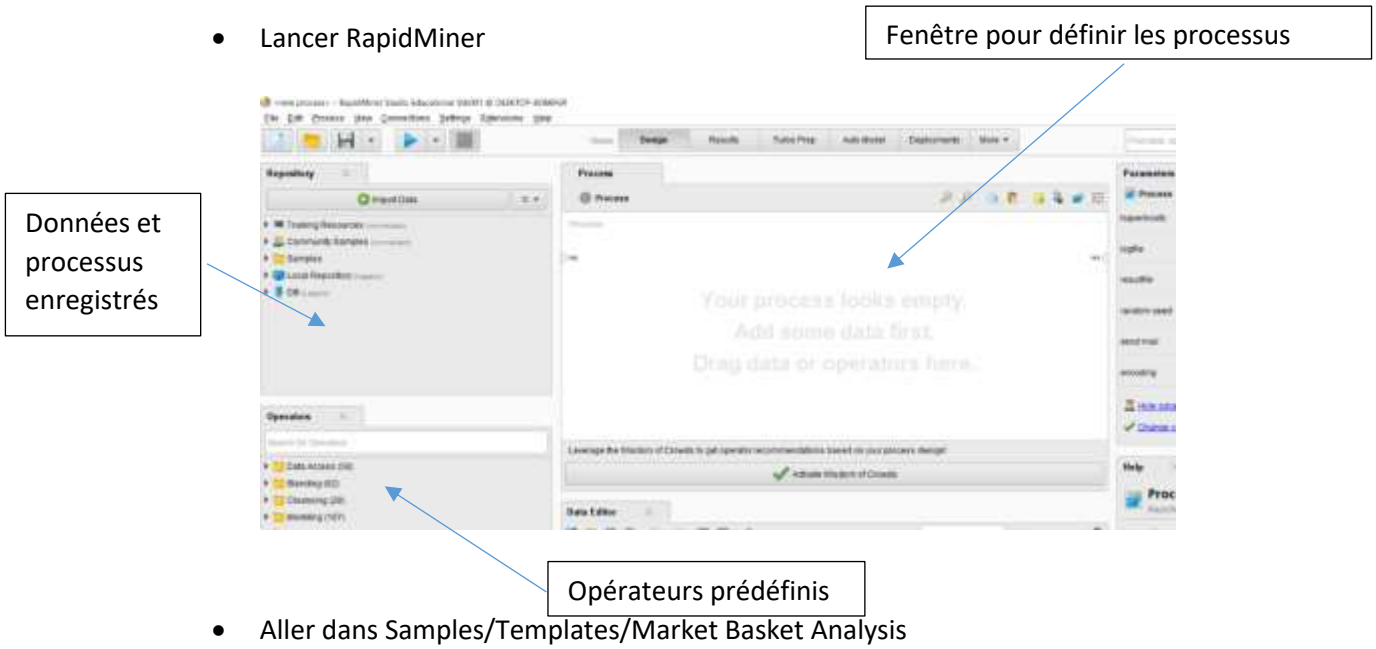
RapidMiner est facile d'utilisation car il est basé sur un ensemble d'opérateurs pré-définis (on peut en ajouter soi-même) qu'il s'agit d'enchainer sous la forme de workflow pour créer un processus d'extraction de connaissances. Le tout, se faisant d'une manière visuelle à base d'opérations « Glisser » et « Relier ».

Pour illustrer, nous allons travailler sur le fichier de données qui s'appelle « Transactions » qui se trouve déjà dans RapidMiner une fois ce dernier installé (on verra comment le retrouver). Il s'agit d'une table de la forme Transactions(Invoice, Product1, Order, Sales Value). Chaque ligne de cette table signifie que dans telle facture, on a un produit qui a été commandé pour telle quantité (Order) et qui a été facturé pour telle montant. Si 2 produits différents sont mentionnés dans une même facture, alors on a 2 enregistrements.

On veut extraire à partir de cette table les produits qui sont régulièrement facturés ensemble (association entre produits facturés). Chaque facture est donc vue comme une transaction.

Avant de pouvoir extraire les ensembles fréquents puis les règles d'association, il faut d'abord « mettre en forme » les données pour les présenter sous un format « acceptable » par RapidMiner. En effet, RapidMiner, et plus précisément son implémentation de FP-Growth, requiert à ce que les transactions soient sous la forme d'une table avec une seule colonne et où chaque enregistrement est une liste d'items séparés par un caractère spécial (par exemple '|')[1]. Pour réaliser cette transformation, nous allons utiliser les outils de RapidMiner. Une fois celle-ci effectuée, on peut appliquer FP-Growth et par la suite on applique au résultat de FP-Growth, l'opérateur de génération des règles. Noter que pour le premier, on peut fixer, entre autres paramètres, le support minimal, et pour le second, la confiance minimale.

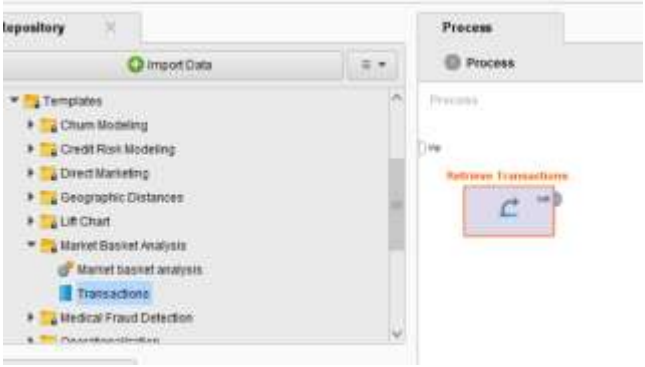
- Lancer RapidMiner



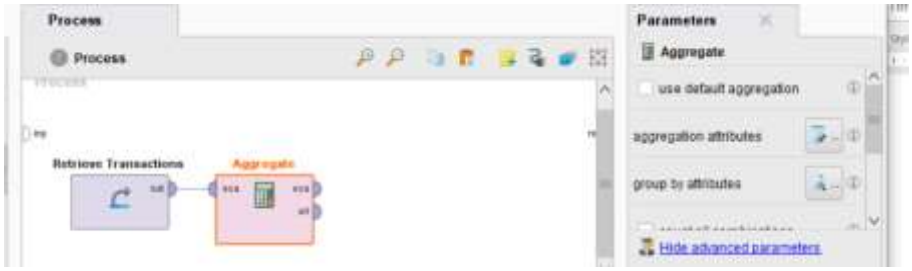
- Aller dans Samples/Templates/Market Basket Analysis
- Double cliquer sur « Transactions » pour voir son contenu.

[1] Il existe deux autres formats possibles mais on ne va considérer pour l'instant que celui-ci.

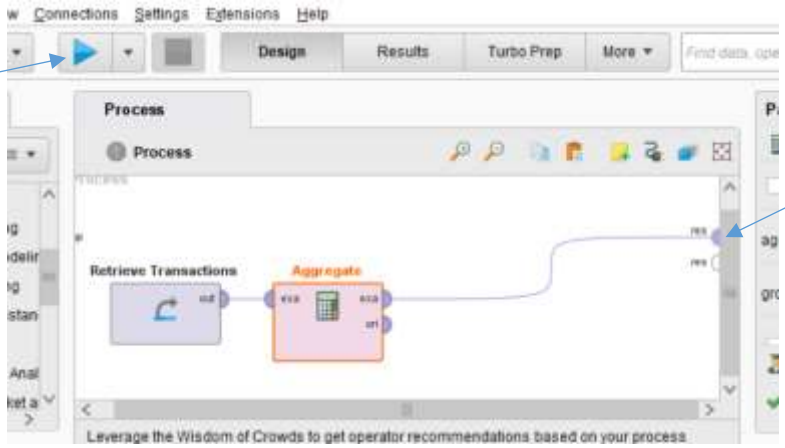
- Glisser « Transactions » vers la fenêtre de création de processus



- Avant de chercher les ensembles fréquents, on agrège les enregistrements de cette table en (i) regroupant ces enregistrements par facture (invoice) et (ii) en concaténant les valeurs de l'attribut Product 1. Pour cela, il faut récupérer l'opérateur d'agrégation. Il suffit de taper « Aggregate » dans la fenêtre des opérateurs puis glisser cet opérateur dans la fenêtre de définition de processus. Faire en sorte que
  - L'entrée d'aggregate soit la table transactions
  - Préciser l'attribut de regroupement (nous c'est Invoice)
  - Préciser l'attribut qu'on veut agréger (c'est Product 1) et la fonction d'agrégation (c'est concatenation)



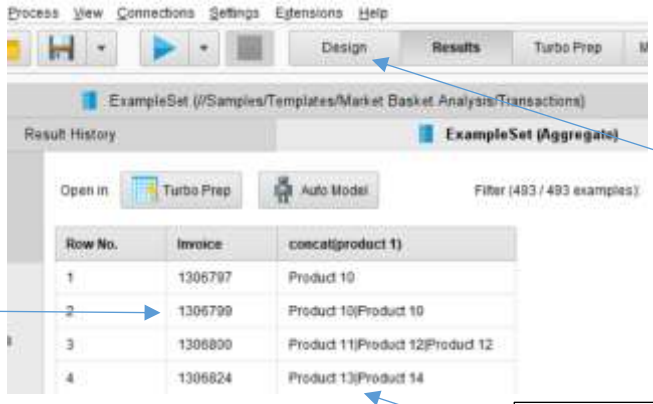
  - On peut déjà chercher à voir le résultat de cette opération : relier la sortie de l'opérateur Aggregate à bouton « res » (pour result) qui se trouve à droite de la fenêtre puis lancer le processus :



Lancer l'exécution

Résultat

  - On obtient le résultat ci-dessous

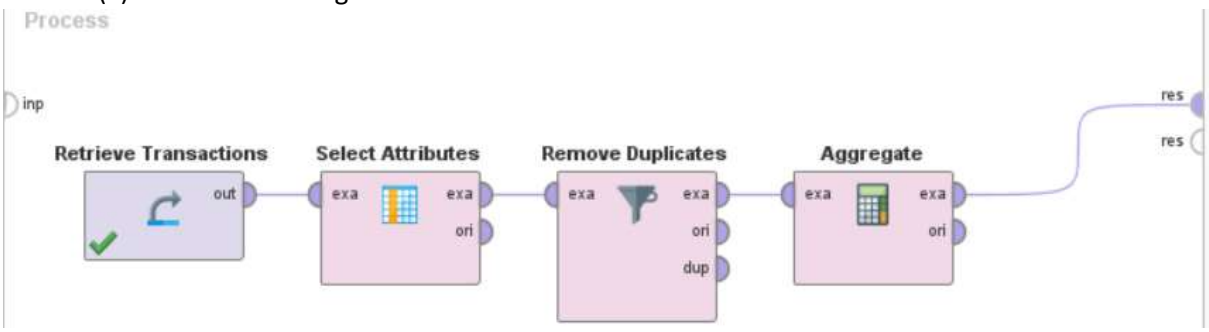


Pour revenir sur la fenêtre de création de processus

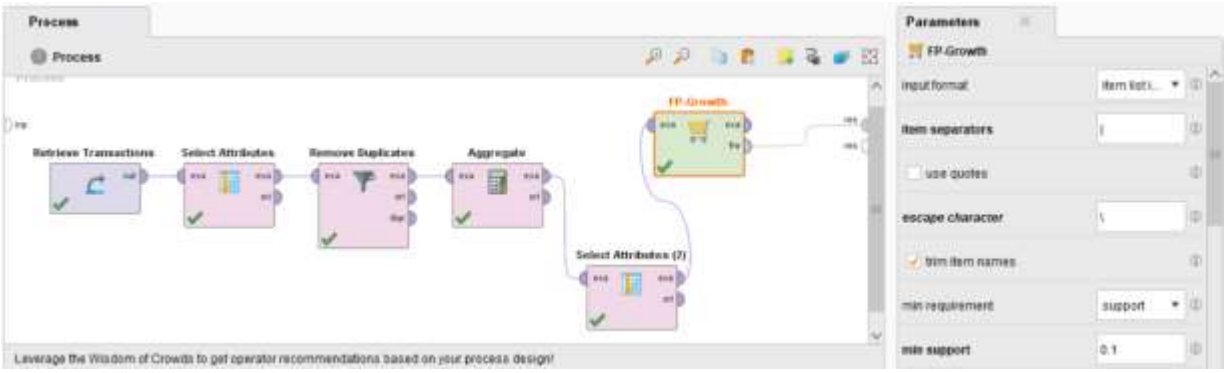
Le produit 10 figure deux fois dans la facture 1306799 !!

Les produits 13 et 14 composent la facture 1306824

- On doit donc éliminer les doublons avant d'agréger. Pour ce faire, on doit (i) sélectionner de la table Transactions juste les deux attributs qui nous intéressent : Invoice et Product 1 et ensuite (ii) éliminer les enregistrements en double.



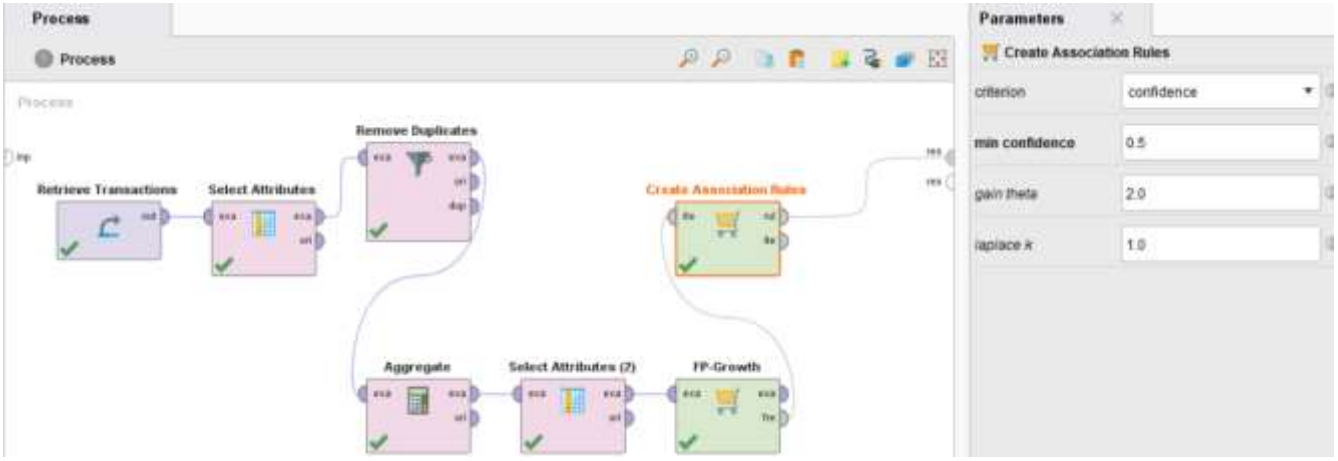
- On peut vérifier que maintenant, chaque produit n'est associé qu'une seule fois à chaque transaction.
- Le résultat de l'opérateur « Aggregate » est une table de transactions à deux colonnes : Invoice et Concat(Product 1). Il faut supprimer la colonne invoice pour qu'elle ait un format accepté par l'opérateur FP-Growth qui trouver les ensembles fréquents (mettons comme support minimal 10%).



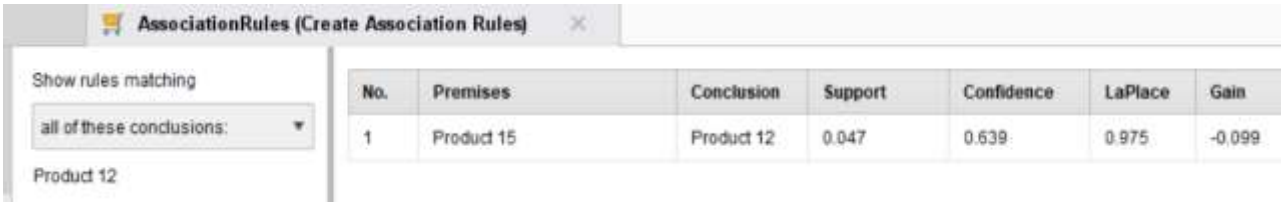
- Le résultat aura la forme suivante :

Size	Support	Item 1	Item 2
1	0.079	Product 23	
1	0.073	Product 15	
1	0.071	Product 26	
1	0.067	Product 13	
1	0.059	Product 21	
1	0.057	Product 24	
1	0.049	Product 19	
1	0.049	product 1	
1	0.047	Product 16	
1	0.043	Product 14	
1	0.037	Product 29	
1	0.028	Product 25	
1	0.028	Product 27	
1	0.024	Product 17	
1	0.024	Product 31	
1	0.022	Product 22	
2	0.034	Product 11	Product 20
2	0.026	Product 12	Product 20

- A partir des ensembles fréquents ainsi obtenus, on peut chercher les règles d'association avec une condition sur la confiance (ex : 50%)



- Le résultat contiendra une seule règle qui est :



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain
1	Product 15	Product 12	0.047	0.639	0.975	-0.099

Exo2 : Application

Soit le fichier Ticket_Clien_Produit(N°Ticket, N°Client, Produit) qui enregistre les produits achetés par un client (disposant d'une carte de fidélité) lors de son passage à la caisse d'un supermarché.

N°Ticket	N°Client	Produit
1	1	A
1	1	B
1	1	C
2	1	A
2	1	C
3	2	A
…	…	…

A partir de cette table, on veut extraire des règles de la forme

Produits → produit

Ces règles peuvent avoir deux sens différents :

1) si tels produits ont été achetés lors d'un passage à la caisse alors tel produit a aussi été acheté lors du même passage, ou bien

2) Si tels produits ont été achetés par un client, alors tel produit a aussi été acheté par le même client, mais pas forcément en même temps.

Utiliser le fichier CSV disponible sur le site du cours pour réaliser ce travail.