# COURS DE DATA MINING
# 7 : MODELISATION NON-SUPERVISEE
# RECHERCHE D'ASSOCIATIONS

**EPF – 4/ 5ème année - Option Ingénierie d'Affaires et de Projets - Finance**

**Bertrand LIAUDET**

**7 : Modélisation non-supervisée - 1 : la recherche d'associations 2**

**Généralités sur la recherche d'associations....2**

Principes 2

Vocabulaire 2

Technique 2

**Algorithme a priori....10**

Principe de l'algorithme 10

Exemple 10

**Induction de règles généralisée : GRI....15**

Principe de l'algorithme 15

**L'analyse des co-occurrences....16**

Principe 16

Exemple 16

**Applications....19**

Le fichier des tickets 19

1ère édition : mai-juin 2006.

2ème édition, revue et corrigée, mai-juin 2007.

3ème édition, revue et corrigée, octobre 2008.

# 7 : MODELISATION NON-SUPERVISEE
# - 1 : LA RECHERCHE D'ASSOCIATIONS

## Généralités sur la recherche d'associations

### Principes

La recherche d'associations consiste à savoir quelles valeurs de variable vont ensemble.

Une association peut se formuler ainsi :

**si telles variables valent tant, alors telles autres variables valent tant.**

C'est une technique descriptive très utilisée dans la grande distribution pour analyser les produits achetés simultanément par un client. Elle permet d'agencer les rayons ou d'organiser les promotions.

Elle est aussi utilisée plus généralement pour agencer les produits packagés : banque, téléphonie, assurance.

Elle peut aussi s'appliquer à n'importe quelles données.

La principale difficulté vient des volumes en jeu : millions de tickets de caisse, dizaines de milliers de produits, conduisant à des milliers de résultats dont beaucoup sont sans intérêt. Dans ces contextes, cette technique demande une grande puissance de calcul.

### Vocabulaire

On parle de :

- Recherche d'associations
- Recherche de règles d'association
- Analyse d'affinité
- Analyse du panier de la ménagère
- Analyse du ticket de caisse

### Technique

#### 1 : Formulation générale, première approche

Une règle d'association a la forme logique générale suivante :

**si telles variables valent tant, alors telles autres variables valent tant.**

si Conditions alors Résultats

$$C \rightarrow R$$

Il s'agit une implication logique, synchronique et non pas d'une relation causale.

Exemple :

si j'achète des fraises alors j'achète des pommes

## 2 : Formulation générale, deuxième approche

L'exemple précédent s'interprète plutôt en disant :

J'achète des fraises **et** des pommes **en mêmes temps**.

Ce qu'on recherche, c'est donc bien les associations, le « et » entre des comportements.

Formellement, il existe une cause, cachée et qui ne nous intéresse pas, qui produit des résultats (des comportements) :

C ⟶□ j'achète des fraises et j'achète des pommes

On a une relation causale, une implication diachronique et non pas une implication logique, synchronique (sur l'implication diachronique et son symbole ⟶□ , cf. cours d'introduction).

En généralisant, on peut écrire :

C ⟶□ R1 et R2 … et Rn

Si C est la cause unique de R1et R2… et Rn, alors on peut en déduire que :

R1 et R2… et Rx → Ry et Rz .. et Rn

C'est à dire plus généralement :

$C \rightarrow R$

## 4 : Indice et seuil de confiance de la règle

Les règles d'association fonctionnent rarement à tous les coups : rarement à 100%. On va donc préciser l'indice de confiance des règles :

**si telles variables valent tant, alors dans X% des cas telles autres variables valent tant.**

X est l'indice de confiance de la règle : IC.

pour une règle du type : $C \rightarrow R$

**IC = probabilité (condition et résultat) / probabilité (condition)**

**IC = p(C,R) / p(C)**

Avec:

p(C,R) = nb (C = vrai et R = vrai) / nb total

p(C) = nb (C = vrai ) / nb total

On arrive à :

**IC = nb (C = vrai et R = vrai) / nb (C = vrai )**

Une règle qui a un indice de confiance trop faible n'est pas intéressante. On sera donc amené à définir un **seuil de confiance.**

## 5 : Indice de support de la règle

Le support de la règle, c'est la population concernée par la règle, donc celle qui vérifie les conditions de la règle.

L'indice de support, c'est probabilité de la condition dans la population d'origine.

pour une règle du type : C → R

**IS = probabilité (condition)**

**IS = p(C)**

Ou encore :

**IS = nb (C = vrai ) / nb total**

- ***Exemple***

Le tableau suivant présente 10 tickets de caisse pour lesquels il y a eu, ou pas, achat de pomme et de fraises.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Pommes** | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Fraises** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |

Calcul de l'indice de confiance de la règle « pomme → fraise » :

p(pomme) = 5/10

p(pomme, fraise) = 4/10

**IC =** (4/10) / (5/10) = 4 / 5 = 80%.

Dans 80% des cas où il y a « pomme », il y a aussi « fraise ».

Calcul de l'indice de support :

**IS =** p(pomme) = 5/10 = 50%.

Donc, 50% des individus sont concernés par cette règle.

## 6 : Les 3 indicateurs de la valeur d'une règle

Il n'existe pas de valeur absolue d'une règle mais seulement des indicateurs statistiques.

- ***IC * IS ou p(C,R***

Le pourcentage de **réussite de la règle dans la population totale** est un indicateur de la valeur d'une règle.

IC * IS donne cette mesure :

**IC * IS = p(C,R)**

L'exemple précédent donne :

**IC * IS =** 80% * 50% = 40%

Cet indicateur montre le taux d'individus (40%) concernés par la règle.

- ***Le « lift »***

Pour qu'une règle soit intéressante, **l'indice de confiance de la règle doit être supérieur à la probabilité absolue du résultat.**

Par exemple, si un résultat à 60% de chance d'arriver dans tous les cas et qu'une règle nous dit que dans tel cas particulier (c'est-à-dire sous telle condition), le résultat à 50% de chance d'arriver, cette règle n'a pas d'intérêt.

Pour qu'une règle soit intéressante, il faut donc que :

**indice de confiance > probabilité (résultat)**

**IC > p(R)**

Le **« lift »,** c'est le taux de progression de la probabilité du résultat grâce à la règle :

**lift = IC / p(R)**

**Un lift intéressant est > 1**

On peut aussi calculer la **progression brute** :

**PB = IC - p(R)**

**Une progression brute intéressante est > 0**

*Exemple*

On reprend le tableau précédent en changeant les valeurs :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pommes | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Fraises | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |

L'indice de confiance de la règle « pomme → fraise » vaut : (4/10) / (5/10) soit 80%.

Donc, dans 80% des cas où il y a « pomme », on trouve aussi « fraise ».

Mais la probabilité des fraises vaut : (9/10) soit 90 %.

Le lift vaut : 80 / 90 = 8/9.

La progression brute vaut : 80 – 90 = -10%.

Lift < 1 et progression brute < 0 : la règle « pomme → fraise » est donc sans intérêt.

*Règle inverse*

Si une règle est inutile (le lift est < 1) alors, la règle inverse est utile :

si C → R est inutile alors C → non(R) est utile

*Exemple*

Dans l'exemple précédent, la règle « pomme → non(fraise) » est utile.

L'indice de confiance de la règle « pomme → non(fraise) » vaut : (1/10) / (5/10) soit 20%.

La probabilité des non(fraise) vaut : (1/10) soit 10%.

Le lift vaut : 20% / 10% = 2.

De 10% de chances on passe à 20 % de chance.

On a 100% (10 +X*10 = 20) de chance en plus de ne pas avoir des fraises quand on a pris des pommes que dans tous les cas.

- ***La capacité de déploiement***

La capacité de déploiement, c'est le pourcentage de ceux qui vérifient les conditions mais pas encore les résultats.

Autrement dit, c'est le pourcentage de la population qui est potentiellement concerné par l'application de la règle.

**Capacité de déploiement = p(C, non R)**

**Exemple :**

On reprend le tableau précédent :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Pommes** | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Fraises** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |

La capacité de déploiement de la règle « pomme → fraise » vaut : (1/10) soit 10%.

## 7 : Type des variables

On a vu dans les exemples que la recherche d'associations se fait sur des variables booléennes.

Cependant, les types continus (numériques) et catégoriels peuvent aussi être pris en compte.

Les types continus doivent d'abord être discrétisés, c'est-à-dire transformés en types catégoriels.

Les types catégoriels sont traités en transformant chaque catégorie (chaque valeur possible pour la variable) en une nouvelle variable qui est donc booléenne.

- ***Exemple***

Soit le tableau suivant :

| Client | Economie | Capital | Revenu |
|---|---|---|---|
| 1 | Moyen | Élevé | Élevé |
| 2 | Faible | Faible | Moyen |
| 3 | Élevé | Moyen | Faible |
| 4 | Moyen | Moyen | Moyen |
| 5 | Faible | Moyen | Très élevé |
| 6 | Élevé | Élevé | Faible |
| 7 | Faible | Faible | Faible |
| 8 | Moyen | Moyen | Élevé |

On va le transformer en matrice creuse :

| Client | Eco = Faible | Eco = Moyen | Eco = Élevé | Cap = Faible | Cap = Moyen | Cap = Élevé | Rev = Faible | Rev = Moyen | Rev = Élevé | Rev = Très élevé |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | 1 | | | | 1 | | | 1 | |
| 2 | 1 | | | 1 | | | | 1 | | |
| 3 | | | 1 | | 1 | | 1 | | | |
| 4 | | 1 | | | 1 | | | 1 | | |
| 5 | 1 | | | | 1 | | | | | 1 |
| 6 | | | 1 | | | 1 | 1 | | | |
| 7 | 1 | | | 1 | | | 1 | | | |
| 8 | | 1 | | | 1 | | | | 1 | |

## 8 : Choix des variables : règle d'association et modèle entité-association

La recherche d'associations traite surtout d'associations entre les produits dans un magasin : son premier domaine d'application est l'analyse du ticket de caisse.

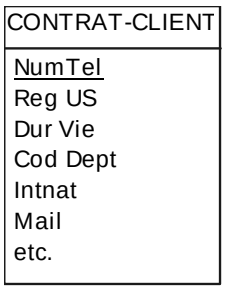
Les variables en jeu sont des variables qui identifient le produit. Dans nos exemples, la variable, c'est le produit lui-même.

On utilise aussi la taxinomie du produit, c'est-à-dire les différents genres dans lesquels on peut classer le produit (par exemple : tel produit est un yaourt, un produit laitier, un dessert, etc.).

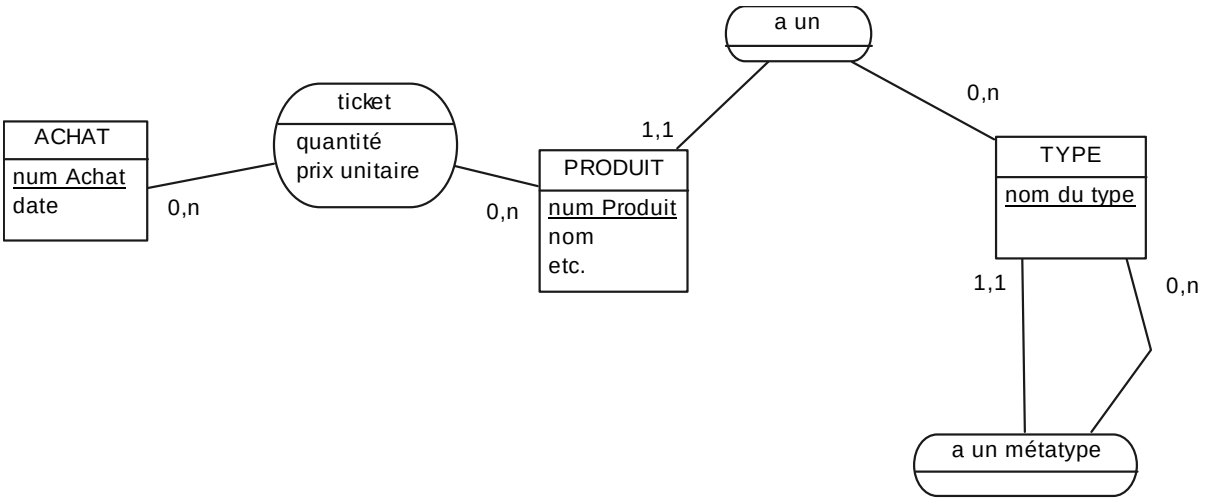
Cette situation est différente de celle de la classification et de l'analyse des données car les attributs avec lesquels on travaille sont souvent des clés étrangères, dans la terminologie du modèle entité-association (MEA).

- ***Exemple :***

La table du fichier d'attrition correspond au MEA suivant :




La table du fichier d'analyse des règles d'association du ticket de caisse correspond au MEA suivant :




Les règles d'association du data mining s'appliquent à l'association non-hiérarchique « ticket ».

Cette association correspond à la table suivante :

Ticket (#numAchat, #numProd, quantité, prix unitaire)

En considérant chaque valeur possible de la variable numProd comme une variable booléenne, on va fabriquer la matrice creuse qui va servir pour la recherche des associations.

On peut aussi travailler au niveau des types ou des méta-types. En faisant la jointure du ticket avec les types, on obtient la table suivante :

Ticket (#numAchat, #numProd, quantité, prix unitaire, #nomType)

En considérant chaque valeur possible de la variable nomType comme une variable booléenne, on peut alors fabriquer la matrice creuse qui va servir pour la recherche des associations.

Dans la table ticket, on trouve par exemple (en considérant nomFruit comme l'équivalent de numProd) :

| **NumTicket** | **nomFruit** |
|---|---|
| 1 | Fraise |
| 1 | Poire |
| 1 | Banane |
| 2 | Citron |
| 2 | Pomme |
| 2 | Banane |
| 3 | Banane |
| Etc. | |

La transformation en matrice creuse donne le résultat suivant :

| numT \ numF | Fraise | Poire | Banane | Citron | Pomme | Etc. |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 1 | 1 | 0 |
| 3 | etc. | etc. | etc. | etc. | etc. | etc. |
| etc. | etc. | etc. | etc. | etc. | etc. | etc. |

A la place des catégories de produit (nom des fruits), on pourrait s'intéresser à des types : fruits exotiques, agrumes, etc.

## 9 : Réduction du nombre de règles

- ***Nombre de règles possibles***

Si on passe en revue toutes les règles possibles pour n attributs, en ne s'intéressant qu'aux règles avec un résultat atomique, on arrive à :

**$n * ( 2^{(n-1)} – 1)$ règles.**

En effet, le nombre de sous-ensembles d'un ensemble de n éléments est égal à : $2^n – 1$.

Or, on a n variables. Pour chaque variable considérée comme une variable cible (le résultat), il reste donc n-1 variables pouvant intervenir dans la condition, soit ( $2^{(n-1)} – 1$ ) n-uplets de variables qui sont des conditions possibles.

On atteint le million de cas à partir de 17 variables et le milliard à partir de 27 variables.

- ***Fixer un seuil de fréquence***

**Fréquence d'une variable booléenne :** nombre de fois où elle vaut 1.

**Probabilité d'une variable booléenne :** fréquence divisée par le nombre total d'individus.

On a la règle suivante de réduction du nombre de règles :

**Si la fréquence d'une variable est trop basse, la variable ne pourra pas être la condition d'une règle d'association.**

Par exemple, s'il n'y a qu'une seule transaction pour laquelle des bananes ont été achetées, on ne pourra pas dire que l'achat de bananes va avec autre chose.

Pour trouver des règles d'association, on commencera donc par **fixer une fréquence minimum pour la recherche** : c'est le **seuil de fréquence**. Ca permet de limiter le nombre de conditions possibles.

## Algorithme a priori

### Principe de l'algorithme

Pour trouver des règles d'association :

1. On commence par fixer une fréquence minimum pour la recherche : un **seuil de fréquence.**
2. **On calcule la fréquence de chaque n-uplet de variables** en partant des n-uplets à une seule variable (singleton), puis en passant à 2, puis à 3, etc.
3. **On réduit le problème** en utilisant la propriété suivante :

   Quel que soit le n-uplet de variables N et quelle que soit la variable V, la fréquence de N est inférieure à celle de N $\cup$ V (N union V).

   On peut donc commencer par éliminer tous les singletons de fréquence inférieure au seuil, et donc tous les n-uplets contenant ces singletons, puis tous les doublons de fréquence inférieure au seuil, et donc tous les n-uplets contenant ces doublons, etc.

   Cette technique permet de réduire considérablement l'ensemble des n-uplets de variables pouvant être la condition d'une règle d'association.

### Exemple

#### Matrice creuse de départ

On reprend le tableau des ventes de fruits en le complétant :

| numT \ numF | Fraise | Poire | Banane | Citron | Pomme | Raisin | Orange |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | | | | |
| 2 | | | 1 | 1 | 1 | | |
| 3 | | | 1 | | 1 | 1 | 1 |
| 4 | | | 1 | | | 1 | 1 |
| 5 | 1 | | | 1 | | | 1 |
| 6 | | | | 1 | 1 | 1 | 1 |
| 7 | | | 1 | | | 1 | |
| 8 | 1 | | | | | 1 | |
| 9 | | | | 1 | 1 | | 1 |
| 10 | | | 1 | | | | 1 |
| 11 | 1 | 1 | | | 1 | | 1 |
| 12 | | | | 1 | 1 | | 1 |
| 13 | | | 1 | 1 | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | | | | 1 |

Tableau n°1 : matrice creuse des transactions

**Détermination des n-uplet candidats pour générer des règles d'association**

On a 7 variables : il y a donc $7*(2^6 - 1) = 441$ n-uplets candidats.

- ***Première réduction du nombre de variables : choix d'un seuil de fréquence***

On fixe un seuil de fréquence : on choisit 4. Ce choix est arbitraire. On considère qu'à moins de 4 / 14 = 29 %, on ne s'intéressera pas au associations.

- ***Calcul de la fréquence pour chaque variable***

Ensuite, on somme les colonnes pour avoir la fréquence par fruit :

| numF / numT | Fraise | Poire | Banane | Citron | Pomme | Raisin | Orange |
|---|---|---|---|---|---|---|---|
| Fréquence | 5 | 3 | 8 | 6 | 7 | 6 | 10 |
| %age | 35,7 % | 21,4 % | 57,1 % | 42,9 % | 50,0 % | 35,7 % | 78,6 % |

Tableau n°2 : fréquence et %age de fréquence par fruit. Singleton candidat pour les règles d'association.

Toutes les variables sont sélectionnées, sauf la Poire.

- ***Calcul de la fréquence pour chaque couple de variables***

Ensuite on croise toutes les variables sélectionnées entre elles et on compte le nombre d'occurrences :

| | **Fraise** | **Banane** | **Citron** | **Pomme** | **Raisin** |
|---|---|---|---|---|---|
| Banane | 2 | | | | |
| Citron | 1 | 2 | | | |
| Pomme | 1 | 2 | **5** | | |
| Raisin | 1 | **4** | 2 | 3 | |
| Orange | 3 | **5** | **5** | **6** | **4** |

Tableau n°3 : couples candidats pour les règles d'association.

On peut supprimer la première ligne (fraise) et la dernière colonne (orange) du tableau.

Il n'y a que 6 couples qui ont une fréquence suffisante.

A noter le rôle prépondérant des oranges.

- ***Calcul de la fréquence pour chaque triplet de variables***

On croise tous les couples trouvés avec tous les singletons.

Toutefois, il n'est pas nécessaire de croiser tous les couples trouvés avec tous les singletons.

En effet, on constate que tous les couples avec fraise sont éliminés : on peut donc retirer la fraise des singletons.

Ensuite, pour une variable donnée, le secondes variables intervenant dans les couples éliminés seront éliminées des singletons. Dans le tableau, pour passer tous les couples en revue, on prend toutes les variables « en ligne et en colonne » (banane-fraise, banane-banane, banane-citron, banane pomme, etc.)

Par exemple, si on prend la variable « banane », on peut éliminer « citron » et « pomme » des singletons pour former un triplet. Il ne reste donc que raisin et orange. Il n'y a donc qu'un seul triplet possible : banane, orange, raisin.

On applique la même méthode à chaque colonne.

| | **Fraise** | **Banane** | **Citron** | **Pomme** | **Raisin** |
|---|---|---|---|---|---|
| Banane | 2 | | | | |
| Citron | 1 | 2 | | | |
| Pomme | 1 | 2 | **5** | | |
| Raisin | 1 | **4** | 2 | 3 | |
| Orange | 3 | **5** | **5** | **6** | **4** |
| **Triplets candidats** | | B-R-O | C-P-O | P-C-O | R-B-O |

Tableau n°4 : couples et triplets candidats pour les règles d'association.

A noter que tous les triplets possibles avec Orange se retrouvent parmi les 4 mis au jour ou sont éliminé du fait qu'un couple du triplet est en deçà du seuil.

On aboutit donc au tableau de triplets suivant :

| **N°** | **Triplets** | **Fréquence** |
|---|---|---|
| 1 | Banane, Raisin, Orange | 3 |
| 2 | Citron, Pomme, Orange | 4 |
| 3 | Pomme, Citron, Orange (vu en 2) | |
| 4 | Raisin, Banane, Orange (vu en 1) | |

Tableau n° 5 : triplets candidats pour les règles d'association

Il n'y a qu'un triplet qui ait la fréquence suffisante. C'est la condition d'arrêt de la fabrication des n-uplet candidats : quand on trouve 0 ou 1 n-uplet, on peut s'arrêter.

On obtient donc 1 triplet en plus des 6 couples, soit 7 n-uplets candidats à la fabrication de règles d'association.

## Génération des règles d'association

- ***Principe de la constitution des règles***

Pour chaque couple (a, b), on peut produire deux règles d'association : a $\rightarrow$ b et b $\rightarrow$ a.

Pour chaque n-uplet, on peut produire n nouvelles règles d'association dont la condition sera constituée de n-1 attributs.

Donc avec 6 couples et 1 triplet, on obtient 6*2 + 3 = 15 règles d'association.

- ***Indice de confiance - IC - et indice de support - IS***

On calcule, pour toutes les règles d'association, les indices de support (IS) et de confiance (IC).

Rappelons que, pour la règle C → R :

**IC = p(C,R) / p(C) = f(C,R) / f(C)**

**IS = p(C) = f(C) / nb total**

p( ) étant la probabilité et f( ) la fréquences ou nombre d'occurrences.

**lift = IC / p(R)** (un lift intéressant est > 1)

**Progression Brute = IC - p(R)** (une progression brute intéressante est > 0)

**Capacité de déploiement = p(C, non R),**

Application à notre exemple :

| **Conditions** | **Résultat** | **Ind. de support** | | **Ind. de conf.** | | **S * C** | | | **Lift** | | **Déploiement** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | f(C) | p(C) | f(C,R) | IC | P(C,R) | f(R) | p(R) | | IC-p(R) | | p(C,nR) |
| Banane | Raisin | 8 | 57,1% | 4 | 50,0 % | 28,6% | 6 | 42,9% | 1,2% | 7,1% | 4 | 28,6% |
| Banane | Orange | 8 | 57,1% | 5 | 62,5 % | 35,7% | 10 | 71,4% | 0,9% | -8,9% | 3 | 21,4% |
| Citron | Pomme | 6 | 42,9% | 5 | 83,3 % | 35,7% | 7 | 50,0% | 1,7% | 33,3% | 1 | 7,1% |
| Citron | Orange | 6 | 42,9% | 5 | 83,3 % | 35,7% | 10 | 71,4% | 1,2% | 11,9% | 1 | 7,1% |
| Pomme | Orange | 7 | 50,0% | 6 | 85,7 % | 42,9% | 10 | 71,4% | 1,2% | 14,3% | 1 | 7,1% |
| Raisin | Orange | 6 | 42,9% | 4 | 66,7 % | 28,6% | 10 | 71,4% | 0,9% | -4,8% | 2 | 14,3% |
| Pomme | Citron | 7 | 50,0% | 5 | 71,4 % | 35,7% | 6 | 42,9% | 1,7% | 28,6% | 2 | 14,3% |
| Raisin | Banane | 6 | 42,9% | 4 | 66,7 % | 28,6% | 8 | 57,1% | 1,2% | 9,5% | 2 | 14,3% |
| Orange | Banane | 10 | 71,4% | 5 | 50,0 % | 35,7% | 8 | 57,1% | 0,9% | -7,1% | 5 | 35,7% |
| Orange | Citron | 10 | 71,4% | 5 | 50,0 % | 35,7% | 6 | 42,9% | 1,2% | 7,1% | 5 | 35,7% |
| Orange | Pomme | 10 | 71,4% | 6 | 60,0 % | 42,9% | 7 | 50,0% | 1,2% | 10,0% | 4 | 28,6% |
| Orange | Raisin | 10 | 71,4% | 4 | 40,0 % | 28,6% | 6 | 42,9% | 0,9% | -2,9% | 6 | 42,9% |
| Citron, Pomme | Orange | 5 | 35,7% | 4 | 80,0 % | 28,6% | 10 | 71,4% | 1,1% | 8,6% | 1 | 7,1% |
| Citron, Orange | Pomme | 5 | 35,7% | 4 | 80,0 % | 28,6% | 7 | 50,0% | 1,6% | 30,0% | 1 | 7,1% |
| Pomme, Orange | Citron | 6 | 42,9% | 4 | 66,7 % | 28,6% | 6 | 42,9% | 1,6% | 23,8% | 2 | 14,3% |

Tableau n°7 : règles d'association avec support et seuil de confiance

Le tableau ci-dessus a été construit avec les données des tableaux n° 1 et 2.

Dans les colonnes « lift », on trouve

Le « lift » : IC / p(R)

La progression brute : IC – p(R )

Dans les colonnes « déploiement », on trouve

f(C, nR) : fréquence des conditions avec la négation du résultat

p(C, nR) : probabilité des conditions avec la négation du résultat

## Interprétation des règles d'association

### ➢ *Classifications*

Pour l'interprétation, on va classer le tableau des règles selon les 3 indicateurs de qualité : par

- **IC * IS** : réussite de la règle dans la population totale

- **lift** : taux de progression de la probabilité du résultat
- **déploiement** : pourcentage de la population concerné par l'application de la règle

On commence par le lift car cet indicateur permet d'éliminer des règles.

➢ ***Classement par lift***

| **Conditions** | **Résultat** | **Ind. de support** | | **Ind. de conf.** | | **IS * IC** | | | **Lift** | | **Déploiement** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F(C) | p(C) | f(C,R) | IC | P(C,R) | f(R) | p(R) | | IC-p(R) | | p(C,nR) |
| Citron | Pomme | 6 | 42,90% | 5 | 83,30% | 35,70% | 7 | 50,00% | 1,70% | 33,30% | 1 | 7,10% |
| Citron, Orange | Pomme | 5 | 35,70% | 4 | 80,00% | 28,60% | 7 | 50,00% | 1,60% | 30,00% | 1 | 7,10% |
| Pomme | Citron | 7 | 50,00% | 5 | 71,40% | 35,70% | 6 | 42,90% | 1,70% | 28,60% | 2 | 14,30% |
| Pomme, Orange | Citron | 6 | 42,90% | 4 | 66,70% | 28,60% | 6 | 42,90% | 1,60% | 23,80% | 2 | 14,30% |
| Pomme | Orange | 7 | 50,00% | 6 | 85,70% | 42,90% | 10 | 71,40% | 1,20% | 14,30% | 1 | 7,10% |
| Citron | Orange | 6 | 42,90% | 5 | 83,30% | 35,70% | 10 | 71,40% | 1,20% | 11,90% | 1 | 7,10% |
| Orange | Pomme | 10 | 71,40% | 6 | 60,00% | 42,90% | 7 | 50,00% | 1,20% | 10,00% | 4 | 28,60% |
| Raisin | Banane | 6 | 42,90% | 4 | 66,70% | 28,60% | 8 | 57,10% | 1,20% | 9,50% | 2 | 14,30% |
| Citron, Pomme | Orange | 5 | 35,70% | 4 | 80,00% | 28,60% | 10 | 71,40% | 1,10% | 8,60% | 1 | 7,10% |
| Orange | Citron | 10 | 71,40% | 5 | 50,00% | 35,70% | 6 | 42,90% | 1,20% | 7,10% | 5 | 35,70% |
| Banane | Raisin | 8 | 57,10% | 4 | 50,00% | 28,60% | 6 | 42,90% | 1,20% | 7,10% | 4 | 28,60% |
| Orange | Raisin | 10 | 71,40% | 4 | 40,00% | 28,60% | 6 | 42,90% | 0,90% | -2,90% | 6 | 42,90% |
| Raisin | Orange | 6 | 42,90% | 4 | 66,70% | 28,60% | 10 | 71,40% | 0,90% | -4,80% | 2 | 14,30% |
| Orange | Banane | 10 | 71,40% | 5 | 50,00% | 35,70% | 8 | 57,10% | 0,90% | -7,10% | 5 | 35,70% |
| Banane | Orange | 8 | 57,10% | 5 | 62,50% | 35,70% | 10 | 71,40% | 0,90% | -8,90% | 3 | 21,40% |

Les « lift » < 1 (ce qui équivaut à une progression brute négative) sont des règles sans intérêt.
Les 7 premières règles se résument à l'association : Citron – Pomme – Orange.
Il ne reste qu'une seule autre association : Raisin – Banane.

➢ ***Classement par IC * IS***

| **Conditions** | **Résultat** | **Ind. de support** | | **Ind. de conf.** | | **IS * IC** | | | **Lift** | | **Déploiement** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F(C) | p(C) | f(C,R) | IC | P(C,R) | f(R) | p(R) | | IC-p(R) | | p(C,nR) |
| Orange | Pomme | 10 | 71,40% | 6 | 60,00% | 42,90% | 7 | 50,00% | 1,20% | 10,00% | 4 | 28,60% |
| Pomme | Orange | 7 | 50,00% | 6 | 85,70% | 42,90% | 10 | 71,40% | 1,20% | 14,30% | 1 | 7,10% |
| Orange | Citron | 10 | 71,40% | 5 | 50,00% | 35,70% | 6 | 42,90% | 1,20% | 7,10% | 5 | 35,70% |
| Orange | Banane | 10 | 71,40% | 5 | 50,00% | 35,70% | 8 | 57,10% | 0,90% | -7,10% | 5 | 35,70% |
| Banane | Orange | 8 | 57,10% | 5 | 62,50% | 35,70% | 10 | 71,40% | 0,90% | -8,90% | 3 | 21,40% |
| Pomme | Citron | 7 | 50,00% | 5 | 71,40% | 35,70% | 6 | 42,90% | 1,70% | 28,60% | 2 | 14,30% |
| Citron | Pomme | 6 | 42,90% | 5 | 83,30% | 35,70% | 7 | 50,00% | 1,70% | 33,30% | 1 | 7,10% |
| Citron | Orange | 6 | 42,90% | 5 | 83,30% | 35,70% | 10 | 71,40% | 1,20% | 11,90% | 1 | 7,10% |
| Orange | Raisin | 10 | 71,40% | 4 | 40,00% | 28,60% | 6 | 42,90% | 0,90% | -2,90% | 6 | 42,90% |
| Banane | Raisin | 8 | 57,10% | 4 | 50,00% | 28,60% | 6 | 42,90% | 1,20% | 7,10% | 4 | 28,60% |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pomme, Orange | Citron | 6 | 42,90% | 4 | 66,70% | 28,60% | 6 | 42,90% | 1,60% | 23,80% | 2 | 14,30% |
| Raisin | Banane | 6 | 42,90% | 4 | 66,70% | 28,60% | 8 | 57,10% | 1,20% | 9,50% | 2 | 14,30% |
| Raisin | Orange | 6 | 42,90% | 4 | 66,70% | 28,60% | 10 | 71,40% | 0,90% | -4,80% | 2 | 14,30% |
| Citron, Orange | Pomme | 5 | 35,70% | 4 | 80,00% | 28,60% | 7 | 50,00% | 1,60% | 30,00% | 1 | 7,10% |
| Citron, Pomme | Orange | 5 | 35,70% | 4 | 80,00% | 28,60% | 10 | 71,40% | 1,10% | 8,60% | 1 | 7,10% |

On obtient les mêmes résultat qu'avec le lift.

- ***Classement par capacité de déploiement***

| **Conditions** | **Résultat** | **Ind. de support** | | **Ind. de conf.** | | **IS * IC** | | | **Lift** | | **Déploiement** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F(C) | P(C) | f(C,R) | IC | P(C,R) | f(R) | p(R) | | IC-p(R) | | p(C,nR) |
| Orange | Raisin | 10 | 71,40% | 4 | 40,00% | 28,60% | 6 | 42,90% | 0,90% | -2,90% | 6 | 42,90% |
| Orange | Citron | 10 | 71,40% | 5 | 50,00% | 35,70% | 6 | 42,90% | 1,20% | 7,10% | 5 | 35,70% |
| Orange | Banane | 10 | 71,40% | 5 | 50,00% | 35,70% | 8 | 57,10% | 0,90% | -7,10% | 5 | 35,70% |
| Orange | Pomme | 10 | 71,40% | 6 | 60,00% | 42,90% | 7 | 50,00% | 1,20% | 10,00% | 4 | 28,60% |
| Banane | Raisin | 8 | 57,10% | 4 | 50,00% | 28,60% | 6 | 42,90% | 1,20% | 7,10% | 4 | 28,60% |
| Banane | Orange | 8 | 57,10% | 5 | 62,50% | 35,70% | 10 | 71,40% | 0,90% | -8,90% | 3 | 21,40% |
| Pomme | Citron | 7 | 50,00% | 5 | 71,40% | 35,70% | 6 | 42,90% | 1,70% | 28,60% | 2 | 14,30% |
| Pomme, Orange | Citron | 6 | 42,90% | 4 | 66,70% | 28,60% | 6 | 42,90% | 1,60% | 23,80% | 2 | 14,30% |
| Raisin | Banane | 6 | 42,90% | 4 | 66,70% | 28,60% | 8 | 57,10% | 1,20% | 9,50% | 2 | 14,30% |
| Raisin | Orange | 6 | 42,90% | 4 | 66,70% | 28,60% | 10 | 71,40% | 0,90% | -4,80% | 2 | 14,30% |
| Citron | Pomme | 6 | 42,90% | 5 | 83,30% | 35,70% | 7 | 50,00% | 1,70% | 33,30% | 1 | 7,10% |
| Citron, Orange | Pomme | 5 | 35,70% | 4 | 80,00% | 28,60% | 7 | 50,00% | 1,60% | 30,00% | 1 | 7,10% |
| Pomme | Orange | 7 | 50,00% | 6 | 85,70% | 42,90% | 10 | 71,40% | 1,20% | 14,30% | 1 | 7,10% |
| Citron | Orange | 6 | 42,90% | 5 | 83,30% | 35,70% | 10 | 71,40% | 1,20% | 11,90% | 1 | 7,10% |
| Citron, Pomme | Orange | 5 | 35,70% | 4 | 80,00% | 28,60% | 10 | 71,40% | 1,10% | 8,60% | 1 | 7,10% |

L'observation des capacités de développement montre que le triplet Citron – Pomme – Orange est intéressant à déployer, tandis que le couple Raisin – Banane l'est moins.

## Induction de règles généralisée : GRI

### Principe de l'algorithme

L'algorithme d'induction de règles généralisée - GRI - permet de traiter des conditions de tous types : numérique, catégorielles et booléennes.

Les résultats doivent toutefois être des données catégorielles.

L'algorithme fonctionne à peu près comme le précédent

## L'analyse des co-occurrences

### Principe

#### Règle avec une seule condition

Une simplification de la recherche des règles d'association consiste à ne s'intéresser qu'aux règles avec une et une seule variable condition et une et une seule variable résultat :

une variable condition → une variable résultat

Autrement dit, on va s'intéresser à toutes les co-occurrences des variables.

#### Type des variables

Cette analyse peut s'appliquer aux variables booléennes, mais aussi aux variables catégorielles et continues à condition de les transformer en variables booléennes. Attention, pour que les résultats soient exploitables, il ne doit pas y avoir trop de catégories.

#### Mesure de la qualité

- La mesure du nombre de co-occurrences correspond à l'indice de support.
- Le classement de l'indice de support pour toutes les couples de variables analysés permet de se faire une première idée de quelles variables vont avec quelles variables.
- Ce n'est qu'une première idée qui pourra être remise en cause par une recherche d'associations du fait de la prise en compte des indicateurs de qualité.

#### Représentation graphique

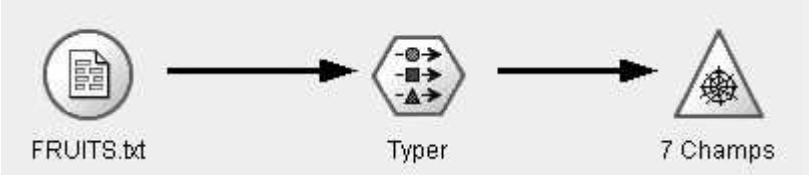
On peut donner une représentation graphique des résultats qui facile l'interprétation (nœud graphique-relation de Clementine).

Une fort nombre de co-occurrence entre deux variables (ou catégories) montre une association entre ces deux variables (ou catégories).

Un fort nombre de co-occurrence pour plusieurs couples de variables (ou catégories) qui partagent une variable (ou catégorie) commune montre un réseau de variables (ou catégories) associées. Un réseau doit être fermé.

### Exemple

#### Le fichier des fruits




Flux Clementine de l'analyse des co-occurrences : nœud graphique / relation

A noter que le nœud « relation » doit être paramétrer pour les booléens vrais uniquement.

On obtient les résultats chiffrés suivants :

**Récapitulatif** | **Commandes**

Liens forts

| Liens | Champ 1 | Champ 2 |
|---|---|---|
| 6 | Pomme = "T" | Orange = "T" |
| 5 | Citron = "T" | Pomme = "T" |
| 5 | Banane = "T" | Orange = "T" |
| 5 | Citron = "T" | Orange = "T" |

Liens moyens

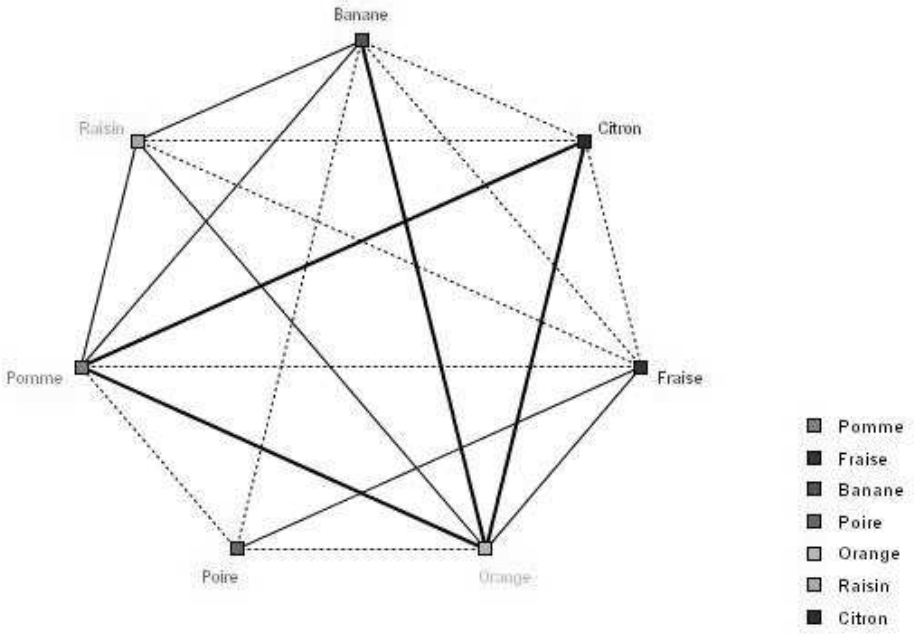
| Liens | Champ 1 | Champ 2 |
|---|---|---|
| 4 | Banane = "T" | Raisin = "T" |
| 4 | Raisin = "T" | Orange = "T" |
| 3 | Fraise = "T" | Poire = "T" |
| 3 | Banane = "T" | Pomme = "T" |
| 3 | Pomme = "T" | Raisin = "T" |
| 3 | Fraise = "T" | Orange = "T" |

Liens faibles

| Liens | Champ 1 | Champ 2 |
|---|---|---|
| 2 | Fraise = "T" | Banane = "T" |
| 2 | Poire = "T" | Banane = "T" |
| 2 | Banane = "T" | Citron = "T" |
| 2 | Citron = "T" | Raisin = "T" |
| 2 | Poire = "T" | Orange = "T" |
| 1 | Fraise = "T" | Citron = "T" |
| 1 | Fraise = "T" | Raisin = "T" |
| 1 | Fraise = "T" | Pomme = "T" |
| 1 | Poire = "T" | Pomme = "T" |

A noter que dans l'onglet « Commandes » on peut choisir les limites pour le nombre d'occurrences de lien forts, moyens et faibles.

Ce tableau donne le réseau graphique suivant :



Sur ce schéma, on constate :

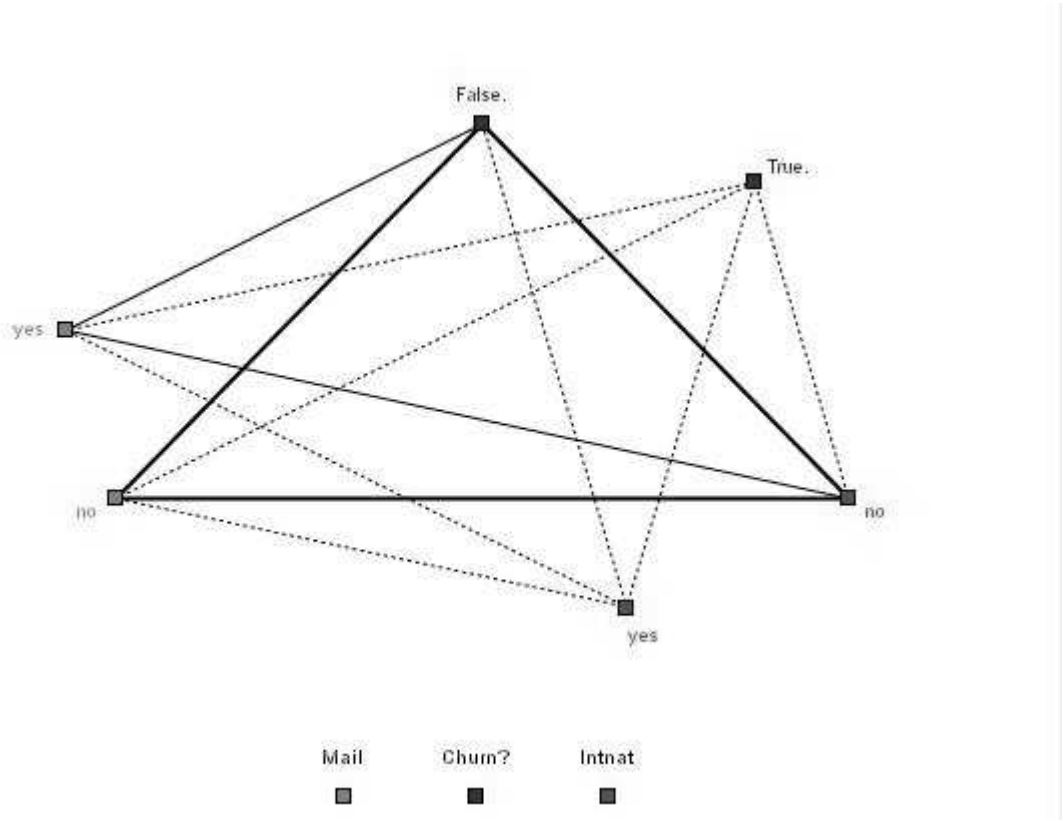
- un réseau fermé entre Citron – Pomme – Orange.
- une association forte entre Orange – Banane
- plusieurs associations moyennes.

Si on compare ces résultats avec ceux de l'algorithme a priori, on constate que :

- On retrouve le triplet Citron – Pomme – Orange.
- On n'a pas mis au jour Raisin – Banane.
- On a gardé Orange – Banane qui avait été éliminé par son lift < 1

On voit donc que l'analyse des co-occurrences est insuffisante pour ce type de recherche.

**Le fichier de churn**




Le réseau ci-dessous nous montre le lien fort entre l'absence d'option international, l'absence de mail et le fait de rester (churn à faux).

On voit aussi que l'option international a un impact plus négatif que l'option mail.

## Applications

### Le fichier des tickets

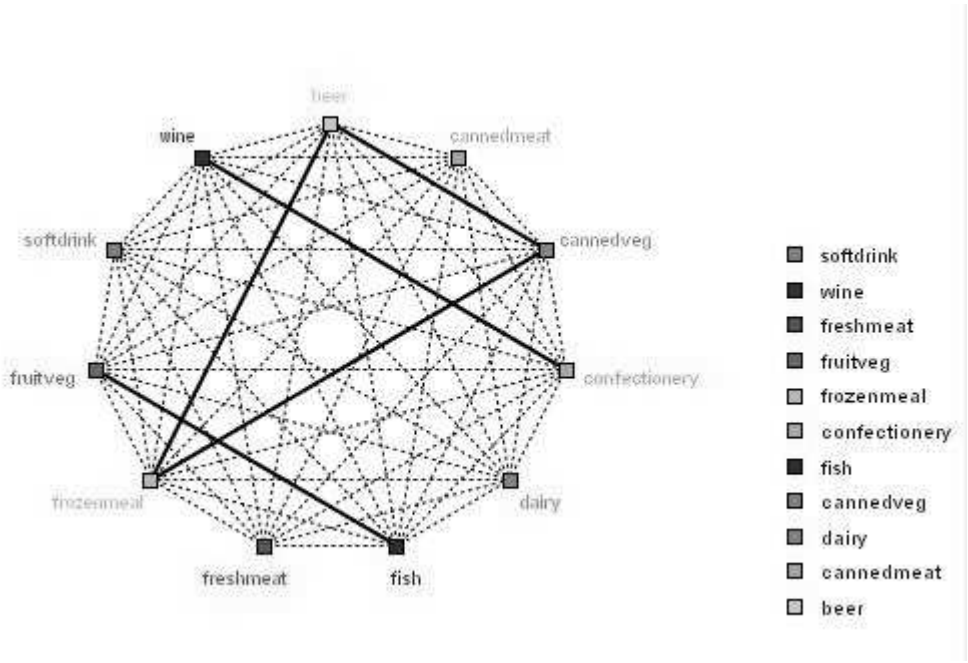
On va appliquer l'algorithme du GRI à un fichier de tickets.

**Dictionnaire des données :**

| Champ | Type | Valeurs |
|---|---|---|
| cardid | Sans type | |
| value | Intervalle | [10.007,49.8863] |
| pmethod | Ensemble | CARD,CASH,CHEQUE |
| sex | Booléen | M/F |
| homeown | Booléen | YES/NO |
| income | Intervalle | [10200,30000] |
| age | Intervalle | [16,50] |
| fruitveg | Booléen | T/F |
| freshmeat | Booléen | T/F |
| dairy | Booléen | T/F |
| cannedveg | Booléen | T/F |
| cannedmeat | Booléen | T/F |
| frozenmeal | Booléen | T/F |
| beer | Booléen | T/F |
| wine | Booléen | T/F |
| softdrink | Booléen | T/F |
| fish | Booléen | T/F |
| confectionery | Booléen | T/F |

Les attributs de « fruitveg » à « confectionery » sont les attributs booléens qui disent si un article a été acheté ou pas.

## Analyse des co-occurrences




Nœud « Graphique / Relations » de Clementine

L'analyse des co-occurrences montre un lien entre :

beer – cannedveg – frozenmeal (réseau fermé)

wine – confectionney

fruitveg - fish

Les résultats chiffrés sont les suivants:

Liens forts

| Liens | Champ 1 | Champ 2 |
|---|---|---|
| 173 | cannedveg = "T" | frozenmeal = "T" |
| 170 | frozenmeal = "T" | beer = "T" |
| 167 | cannedveg = "T" | beer = "T" |
| 145 | fruitveg = "T" | fish = "T" |
| 144 | wine = "T" | confectionery = "T" |

Liens moyens

Liens faibles

| Liens | Champ 1 | Champ 2 |
|---|---|---|
| 97 | cannedveg = "T" | wine = "T" |
| 90 | frozenmeal = "T" | fish = "T" |
| 89 | cannedveg = "T" | fish = "T" |
| 89 | fruitveg = "T" | beer = "T" |
| 86 | fruitveg = "T" | frozenmeal = "T" |
| 86 | fruitveg = "T" | cannedveg = "T" |
| 86 | fish = "T" | confectionery = "T" |
| 85 | beer = "T" | fish = "T" |
| 84 | fruitveg = "T" | wine = "T" |
| 82 | fruitveg = "T" | confectionery = "T" |

Nœud « Graphique / Relations » de Clementine

## Induction de règles généralisée

| Résultat | Conditions (Antécédents) | n° règle | nb cond. | P(C) | P(C,R) / P(C) | P(C,R) | Lift | Déploy. |
|---|---|---|---|---|---|---|---|---|
| cannedveg | freshmeat and frozenmeal and beer | 14 | 30 | 3,00 | 96,67 | 2,90 | 3,19 | 0,10 |
| frozenmeal | freshmeat and cannedveg and beer | 16 | 31 | 3,10 | 93,55 | 2,90 | 3,10 | 0,21 |
| cannedveg | cannedmeat and frozenmeal and beer | 13 | 40 | 4,00 | 90,00 | 3,60 | 2,97 | 0,44 |
| Beer | fruitveg and cannedveg and frozenmeal | 9 | 45 | 4,50 | 88,89 | 4,00 | 3,03 | 0,56 |
| Beer | freshmeat and cannedveg and frozenmeal | 17 | 33 | 3,30 | 87,88 | 2,90 | 3,00 | 0,46 |
| frozenmeal | cannedveg and beer | 1 | 167 | 16,70 | 87,43 | 14,60 | 2,90 | 2,40 |
| frozenmeal | fruitveg and cannedveg and beer | 12 | 46 | 4,60 | 86,96 | 4,00 | 2,88 | 0,69 |
| Beer | dairy and cannedveg and frozenmeal | 19 | 23 | 2,30 | 86,96 | 2,00 | 2,97 | 0,34 |
| frozenmeal | dairy and cannedveg and beer | 20 | 23 | 2,30 | 86,96 | 2,00 | 2,88 | 0,34 |
| cannedveg | frozenmeal and beer | 2 | 170 | 17,00 | 85,88 | 14,60 | 2,83 | 2,80 |
| cannedveg | fruitveg and frozenmeal and beer | 15 | 47 | 4,70 | 85,11 | 4,00 | 2,81 | 0,82 |
| Beer | cannedveg and frozenmeal | 3 | 173 | 17,30 | 84,39 | 14,60 | 2,88 | 3,20 |
| cannedveg | cannedmeat and beer | 18 | 60 | 6,00 | 68,33 | 4,10 | 2,26 | 2,78 |
| frozenmeal | Beer | 4 | 293 | 29,30 | 58,02 | 17,00 | 1,92 | 21,20 |
| cannedveg | Frozenmeal | 6 | 302 | 30,20 | 57,28 | 17,30 | 1,89 | 22,52 |
| frozenmeal | Cannedveg | 7 | 303 | 30,30 | 57,10 | 17,30 | 1,89 | 22,76 |
| Beer | Frozenmeal | 5 | 302 | 30,20 | 56,29 | 17,00 | 1,92 | 23,45 |
| Beer | Cannedveg | 8 | 303 | 30,30 | 55,12 | 16,70 | 1,88 | 24,67 |
| Wine | Confectionery | 10 | 276 | 27,60 | 52,17 | 14,40 | 1,82 | 25,30 |
| confectionery | Wine | 11 | 287 | 28,70 | 50,17 | 14,40 | 1,82 | 28,51 |

**Bilan du nœud « Modélisation – GRI » de Clementine**

- ***Légende***

- Le n° de la règle est un numéro de référence.
- Nb anté. est le nombre d'occurrences des conditions : nb(C)
- p(C) : c'est la probabilité des conditions. On l'appelle aussi indice de support de la règle (IS) ou indice de prise en charge de la règle.
- p(C,R) : probabilité des conditions et du résultat ensembles. Vaut : IS * IC.
- p(C,R) / p(C) : c'est l'indice de confiance de la règle (IC)
- Lift = IC / p(R)
- Deploy est la capacité de déploiement : p(C, non R).

- ***Interprétation et conclusions***

Pour interpréter le tableau précédent, il faut le trier soit par IC*IS, soit par lift, soit par capacité de déploiement.

| Résultat | Conditions (Antécédents) | n° règle | nb Anté. | IS | IC | IS * IC | Lift | Déploy. |
|---|---|---|---|---|---|---|---|---|
| confectionery | Wine | 11 | 287 | 28,70 | 50,17 | 14,40 | 1,82 | 28,51 |
| Wine | Confectionery | 10 | 276 | 27,60 | 52,17 | 14,40 | 1,82 | 25,30 |
| Beer | Cannedveg | 8 | 303 | 30,30 | 55,12 | 16,70 | 1,88 | 24,67 |
| Beer | Frozenmeal | 5 | 302 | 30,20 | 56,29 | 17,00 | 1,92 | 23,45 |
| Frozenmeal | Cannedveg | 7 | 303 | 30,30 | 57,10 | 17,30 | 1,89 | 22,76 |
| Cannedveg | Frozenmeal | 6 | 302 | 30,20 | 57,28 | 17,30 | 1,89 | 22,52 |
| frozenmeal | Beer | 4 | 293 | 29,30 | 58,02 | 17,00 | 1,92 | 21,20 |
| Beer | cannedveg and frozenmeal | 3 | 173 | 17,30 | 84,39 | 14,60 | 2,88 | 3,20 |
| Cannedveg | frozenmeal and beer | 2 | 170 | 17,00 | 85,88 | 14,60 | 2,83 | 2,80 |
| Cannedveg | cannedmeat and beer | 18 | 60 | 6,00 | 68,33 | 4,10 | 2,26 | 2,78 |
| frozenmeal | cannedveg and beer | 1 | 167 | 16,70 | 87,43 | 14,60 | 2,90 | 2,40 |
| Cannedveg | Fruitveg and frozenmeal and beer | 15 | 47 | 4,70 | 85,11 | 4,00 | 2,81 | 0,82 |
| frozenmeal | Fruitveg and cannedveg and beer | 12 | 46 | 4,60 | 86,96 | 4,00 | 2,88 | 0,69 |
| Beer | Fruitveg and cannedveg and frozenmeal | 9 | 45 | 4,50 | 88,89 | 4,00 | 3,03 | 0,56 |
| Beer | freshmeat and cannedveg and frozenmeal | 17 | 33 | 3,30 | 87,88 | 2,90 | 3,00 | 0,46 |
| Cannedveg | cannedmeat and frozenmeal and beer | 13 | 40 | 4,00 | 90,00 | 3,60 | 2,97 | 0,44 |
| Beer | Dairy and cannedveg and frozenmeal | 19 | 23 | 2,30 | 86,96 | 2,00 | 2,97 | 0,34 |
| frozenmeal | Dairy and cannedveg and beer | 20 | 23 | 2,30 | 86,96 | 2,00 | 2,88 | 0,34 |
| frozenmeal | freshmeat and cannedveg and beer | 16 | 31 | 3,10 | 93,55 | 2,90 | 3,10 | 0,21 |
| Cannedveg | freshmeat and frozenmeal and beer | 14 | 30 | 3,00 | 96,67 | 2,90 | 3,19 | 0,10 |

**Bilan du nœud « Modélisation – GRI » de Clementine, trié par « déploy »**

On voit qu'on a 7 règles avec une capacité de déploiement très supérieure à celles des autres. Ces règles sont les règles avec une seule condition.