



## République Algérienne Démocratique et Populaire



Institut national des télécommunications et des technologies de l'information  
et de la communication  
VEON OTA Djezzy



# Rapport de Stage : Product Affinity

Par :

# KERKOURI Mohamed Amine

**Encadrer Par :**

**Fethi ABDLRAHMANE**

**Mohamed SELAMA**

*Alger, le : 31/07/2018*

<i>Chapitre</i>	<i>Page</i>
<i>Table de Matière</i>	<b>2</b>
<i>1. Présentation du stage et de l'entreprise</i>	<b>3</b>
<i>2. Evolution des Data dans le monde</i>	<b>4</b>
<i>2.1 Définition des Data</i>	<b>4</b>
<i>2.2. Importance d'axer sur les données pour les Business</i>	<b>4</b>
<i>3. Data Mining (Exploration de données)</i>	<b>5</b>
<i>3.1. Les taches principales du Data Mining</i>	<b>6</b>
<i>4. Machine Learning</i>	<b>7</b>
<i>4.1. Définition</i>	<b>7</b>
<i>4.1 Algorithme d'apprentissage supervisé (SVM)</i>	<b>8</b>
<i>4.2. Algorithme d'apprentissage non supervisé (K-MEANS)</i>	<b>9</b>
<i>5. Association Rules</i>	<b>10</b>
<i>5.1. Définition</i>	<b>10</b>
<i>5.2. Notion AR</i>	<b>11</b>
<i>5.3.1. Algorithme Apriori</i>	<b>12</b>
<i>5.3.2. Algorithme ECLAT</i>	<b>13</b>
<i>5.3.3. Algorithme FP-Growth</i>	<b>15</b>
<i>5.4. Comparaison des Algorithme</i>	<b>16</b>
<i>6. Product and Business Affinity</i>	<b>17</b>
<i>6.1. L'Up-Selling, Down-Selling, Cross-Selling</i>	<b>18</b>
<i>7. Le cas d'utilisation du de Product Affinity au sein de Djezzy</i>	<b>18</b>
<i>8. Conclusion</i>	<b>21</b>
<i>Bibliographie</i>	<b>22</b>

## 1. Présentation du stage et de l'entreprise :

Djezzy (en arabe : جازى?), officiellement Optimum Telecom Algérie S.p.a (ou simplement OTA) et anciennement Orascom Telecom Algérie S.p.a, est un opérateur de téléphonie mobile algérien qui a été créé le 11 juillet 2001 avant d'ouvrir son réseau en février 2002. Leader des technologies de communication mobiles avec plus de 17 millions d'abonnés au mois de décembre 2015, l'entreprise fournit une vaste gamme de services tels que le prépayé, le post-payé, Internet ainsi que les services à valeur ajoutée et le service universel de télécommunication (SUT).

En janvier 2015, le Fonds national d'investissement (FNI) prend le contrôle de 51 % du capital de la société après trois ans de négociation et plus de quatre ans d'activité très réduite. Toutefois, selon les termes de l'accord, le groupe VEON, garde la responsabilité du management de l'entreprise, avec 49 % des actions<sup>4</sup>. Elle compte plus de 4 000 employés.

J'ai intégré l'équipe Analytiques & Data science (sous la responsabilité de Manager Yacine Lahach) au sein de service Data Management, ce dernier se divise en 3 entités, Data waharhouse, Big Data (DMP), Analytiques & Data science.

Etre sous la responsabilité de Mr Fethi Abderrahmane et Mr Mohamed SELAMA ma faciliter d'entrer en contact direct avec des membres de tous les équipes du service et ma facilité mon travail, surtout pour la collecte de données, les ingénieurs de DMP et Data Waharhouse était à très collaboratifs.

Cela m'a permis d'acquérir une vue global sur le système d'information de l'entreprise d'une manière général et le secteur d'activité de télécom en particulier.

Ma participation sur le projet « Product Affinity » a vraiment enrichi mes connaissances sur le métier de télécom (comprendre le comportement des clients ainsi que le suivi de consommation des offres et promotions marketing) et d'acquérir des nouvelles compétences sur le domaine de data science (l'approche de non supervisé en particulier).

Le rapport ci-après présentera la multitude des connaissances et compétences que j'ai acquises pendant mon stage au sein de l'entreprise.

Djezzy	
	
Logo de Djezzy depuis avril 2015.	
Création	11 juillet 2001
Dates clés	2002 : Lancement officiel du réseau Djezzy. 2003 : Le réseau est disponible dans les 48 wilayas. 2004 : Lancement d'AliO OTA. 2014 : Lancement de la 3G. 2016 : Lancement de la 4G.
Forme juridique	SPA <sup>1</sup>
Slogan	Avec elle, tu peux ! (en arabe : معاها تقدر !)
Siège social	Dar El Beida
Direction	Vincenzo Nespoli (président exécutif) Mathieu Galanti (directeur général)
Actionnaires	État algérien (51 %) VimpelCom (49 %)
Activité	Opérateur de télécommunications
Produits	Go, Good, Line, Play, Liberty, Millennium, Imtiaz, AliO OTA, OTAxphone, @migo, Speed
Société mère	Global Telecom Holding VimpelCom
Sociétés soeurs	Beeline, Kyivstar, Wind, Moobilink, BanglaLink
Effectif	+4 000
Site web	<a href="http://www.djezzy.dz/">www.djezzy.dz/</a> (archive)
Chiffre d'affaires	1 796,323 millions \$ (2012) <sup>2</sup>

[modifier](#) • [modifier le code](#) • [voir wikipedia](#)



---

## *2. Evolution des Data dans le monde*

### *2.1 Définition des Data*

***“Facts And Statistics Collected Together For Reference Or Analysis.”***

Les données sont l'ensemble d'informations issues des résultats d'observations ou d'expériences faites délibérément ou à l'occasion d'autres tâches et soumis aux méthodes statistiques.

### *2.2 Importance d'accès aux données pour les Business*

La plupart des entreprises ne capturent qu'une fraction de la valeur potentielle des données et des analytiques, transformer un monde plein de données en un monde axé sur les données est une idée que de nombreuses entreprises ont trouvé des difficultés de les mettre en pratique, il est possible de résumer les données en 4 types de données :

- **Les données d'état**

Ce sont les données logiquement les plus répandues. Elles permettent de mettre en place une base de référence et serviront de plus en plus comme matière première pour alimenter les moteurs d'algorithme des solutions de Big Data, et réaliser du prévisionnel sur le long terme.

- **Les données de localisation**

Extension logique du GPS, ces données se complètent : le GPS fonctionne bien en déplacement, à l'extérieur, mal sur le statique, sur des déplacements courts et surtout en intérieur. Le potentiel est énorme, certes dans la chaîne logistique qui devrait être la première à l'industrialiser, mais également avec un énorme marché grand public, celui de la localisation d'un objet ou d'une personne. Des fonctionnalités qui demandent à bénéficier d'un traitement en temps réel.

- **Les données personnalisées**

Les acteurs du marché sont très prudents dans ce domaine : ils distinguent les données anonymes sur les usages et les préférences individuelles aux données personnelles associées à la vie privée. Toute la difficulté est de pouvoir associer des règles à des usages en passant de la moyenne aux pratiques de l'individu, sans heurter le respect de la vie privée

Les nouvelles règles de protection des données, la loi RGPD qui est rentrée en vigueur le 25 mai 2018 dans ...: l'Europe désormais protégé mieux la vie privée de ses citoyens, cette loi va ne pas tarder pour rentrer en vigueur aussi en Algérie.

- Les données décisionnelles

Principalement associée à l'exploitation des données d'état, mais également aux deux suivantes, les données décisionnelles doivent accompagner la prise de décision, qu'elle soit automatisée ou personnelle. Elles ont donc deux états, l'automatisation et la persuasion.

### 3. Data Mining (Exploration de données)

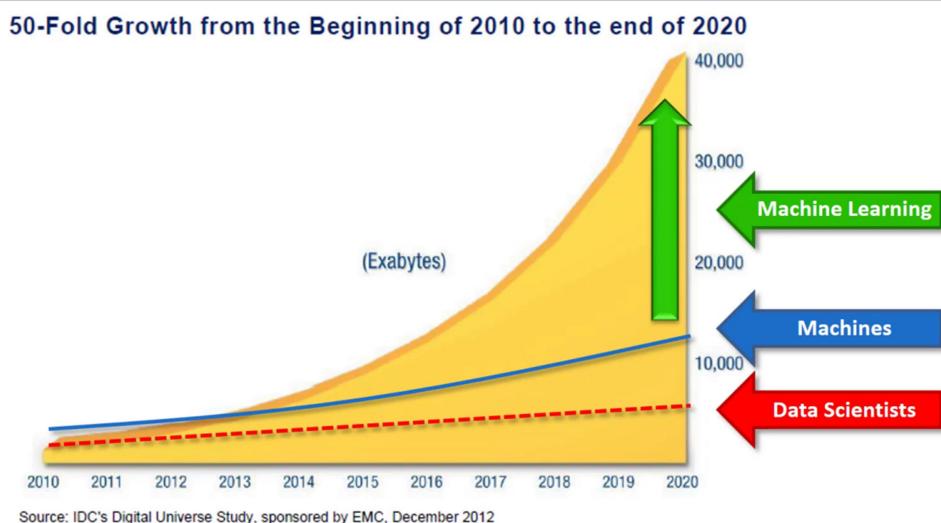


Figure 1 explosion de la quantité des données

En raison de la grande disponibilité d'énormes quantités de données dans une forme électronique, et de la nécessité imminente de transformer ces données en informations et connaissances utiles pour de larges applications comme le Market Analysis, Business Management et Décision Support, le Data Mining a attiré beaucoup d'attention de la part des industriels d'information ces dernières années .

Le Data Mining est défini par :

***“Le processus d'extraction d'informations utiles souvent caché à partir de quantités massives de données complexes.”***

***“L'extraction de connaissances à partir des données est un processus non trivial d'identification de structures inconnues, valides et potentiellement exploitables dans les bases de données.”***

---

*“The Process of discovering interesting knowledge, such as patterns, association ,change, anomalies and significant structures, from large amounts of data stored in databases ,data warehouses, or other information repositories ”*

Le data Mining peut être classifié en deux catégories principales : le data Mining descriptif et Le Data Mining prédictif, Le Premier sert à décrire l'information se trouvant dans les data set alors que le data Mining Prédictif sert à appliquer des modèles inférentiels qui peuvent prédire le comportement des nouvelles données, en se servant des algorithmes intelligents de Machine Learning.

### **3.1 Les taches principales du Data Mining :**

- **Description des classes** : c'est le fait de fournir des résumés concis claire, et précis sur une collection de donnée pour la distinguer des autres classes, parmi les paramètres utilisés dans ce genre de tache : la moyenne, le médian, les sommes, variance, les quartiles, etc....
- **Association** : c'est la découverte de relation d'association et corrélation entre un ensemble d'articles qui se manifeste ensemble dans le data set, par exemple : l'achat d'un produit qui mène à l'achat d'un autre, l'augmentation de primes pour un certain groupe des travailleurs peut munir une meilleur performance alors qu'il la dégrade pour une autre, cela va explorer la relation entre le type du travail et l'effet de l'augmentation de salaire ou prime.
- **Classification** : c'est le découpage des dataset vers plusieurs groupe ou catégorie appelés classes, bien prédéfini, et les libellé. La classification est une catégorie majeure d'algorithme de Machine Learning comme Decision Tree, Logistic Regression, SVM, Naive Bayes.
- **Prédiction et Régression** : c'est la prédiction de valeur pour une donnée non cité dans le dataset, par exemple le prix pour estimer une offre pour une entreprise à partir des offres plus anciennes
- **Clustering** : c'est le découpage des data set vers plusieurs groupes ou catégories, sans qu'ils soient libellée, Le but c'est pouvoir explorer et classifier des donnée dans un contexte plus ou moins inconnue. Une partie majeure des algorithmes de machine learning non supervisé est utilisés
- **Analytiques des Time-series** : c'est l'analyse des data set de phénomène qui sont corrélés avec le temps, et prédire des résultats futures pour les business ou anciennes pour la recherche historique. Un exemple pour cela est le Stock Price prédiction dans la bourse.  
Un cas d'utilité pour cela et définir le meilleur moment pour lancer un nouvel offre téléphonique.

## 4. Machine Learning

### 4.1 Definition

“Optimizing a performance criterion using example data and past experience”.

--Andrew Ng --

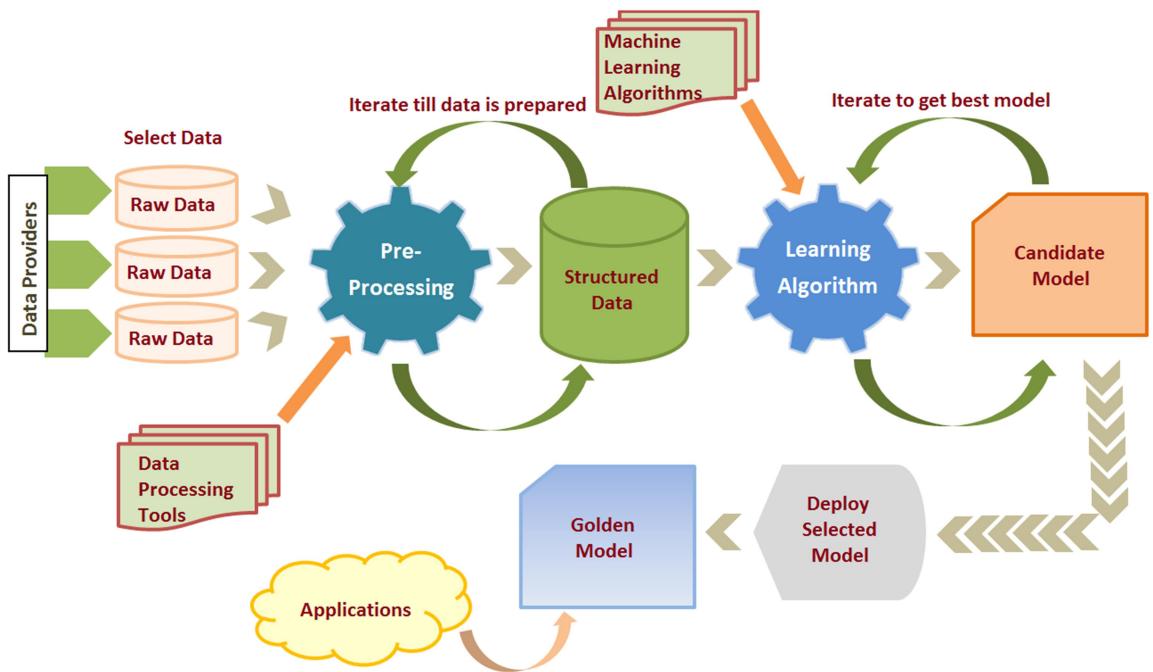


Figure 2 Processus de Machine Learning

Le machine Learning ou **apprentissage pour machine** est la technique moderne pour définir des systèmes qui **apprennent** à résoudre les problèmes sans programmation explicite, la façon dont cette tâche est réalisée, est bien d'utiliser des algorithmes qui puissent trouver les paramètres importants pour cela et puis modélisent le phénomène.

Le Machine Learning est classé en 2 catégories principales :

- **Apprentissage supervisé** : c'est un type qui utilise des data set (appelée training set) pour construire un modèle mathématique puis généralise les règles du data set pour la prédiction.  
Ce genre est découpé lui aussi en algorithmes de classification et algorithmes de régression, le premier est utilisé quand l'espace de résultats est discret, l'autre quand l'espace de prédiction est continue Le Prix par exemple .

- **Apprentissage non supervisé** : c'est une méthode pour l'exploration des données sans avoir un training set. Les algorithmes les plus fameux sont ceux d'Association Rules et de Clustering.

#### 4.1 Algorithme d'apprentissage supervisé :

##### *Support Vector Machine (SVM) :*

Support Vector Machine (SVM) est un algorithme d'apprentissage automatique qui analyse les données pour une analyse de classification et de régression. SVM est une méthode d'apprentissage supervisé qui examine les données et les classe en deux catégories. Un SVM génère une carte des données triées avec les marges entre les deux aussi éloignées que possible. Les SVM sont utilisés dans la catégorisation de texte, la classification d'image, la reconnaissance d'écriture manuscrite et dans les sciences

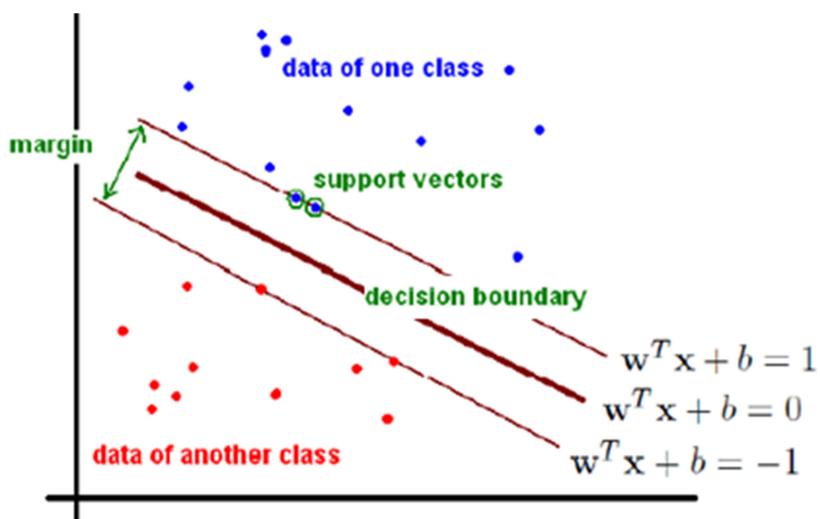


Figure 3 Séparation SVM linéaire

Le SVM est séparateur linéaire ça veut dire qu'il sépare les classes utilisant des lignes droite dans l'espace 2D ou des Hyperplans dans le les diamantions supérieure mais dans le cas où les classes ne sont pas linéairement séparable on utilise la technique de des noyaux (kernels), cette technique consiste à projeter des données dans un espace de dimension supérieure dans lequel le

problème devient linéaire. La projection est contrôlée par les fonctions de kernel comme : RBF, Polynomial kernel, sigmoid kernel.

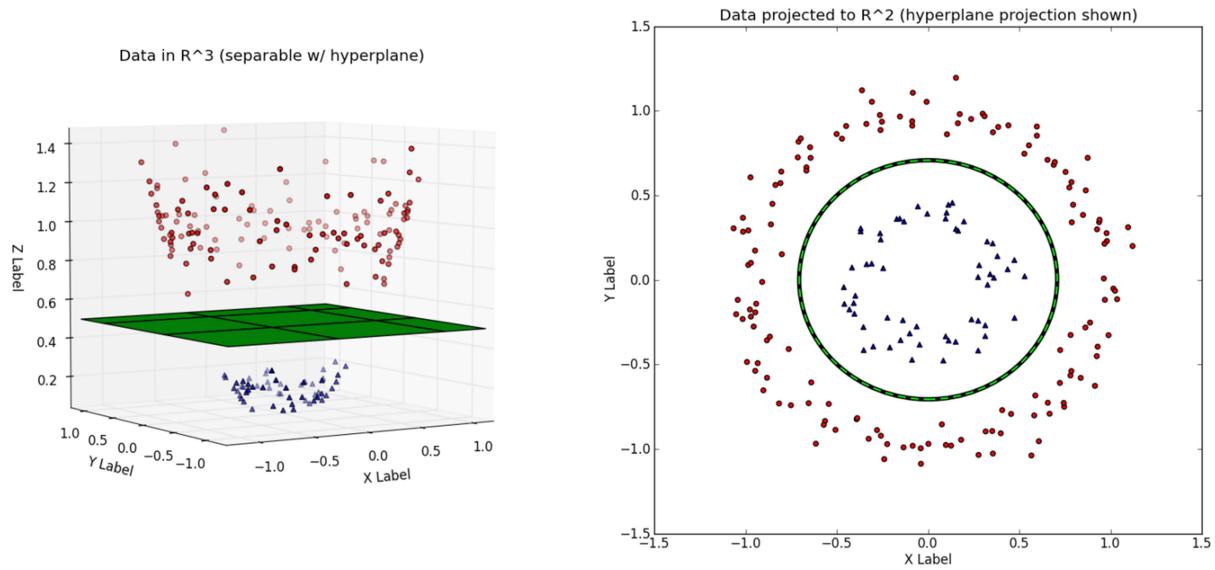
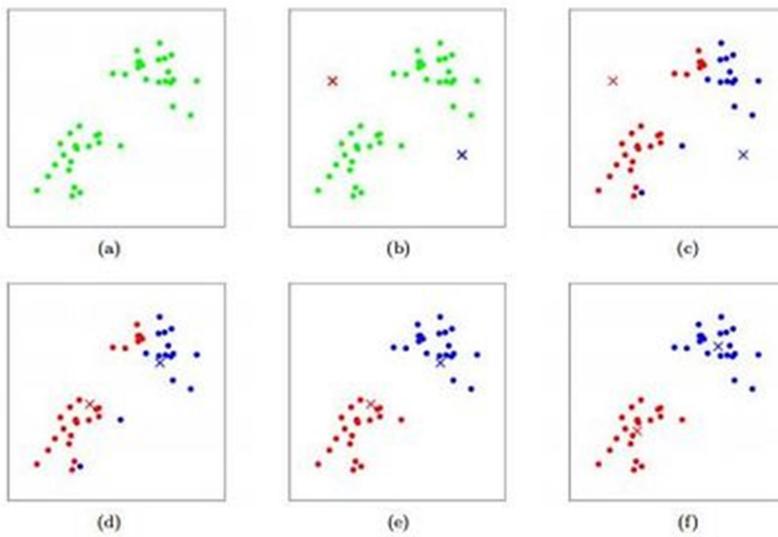


Figure 4 Séparation utilisant un Kernel

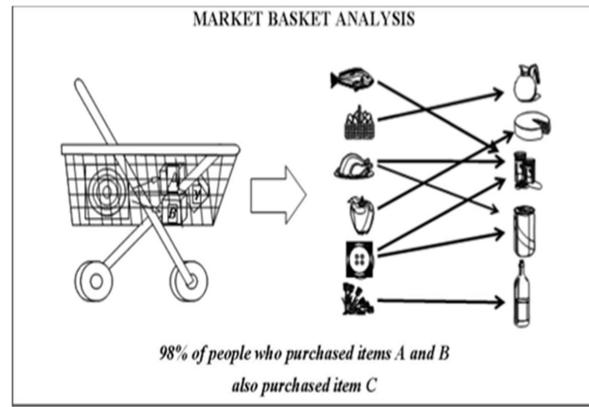
#### 4.2 Algorithme d'apprentissage non supervisé :

##### K-MEANS

Le partitionnement en ***k*-moyennes** (ou ***k-means*** en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier  $k$ , le problème est de diviser les points en  $k$  groupes, souvent appelés *clusters*, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.



L'algorithme choisie K points aléatoire pour K cluster qui vont être transformé en barycentre dans l'espace du data set, puis il calcule la distance entre chaque point du data set et les barycentre, le point est mis en même groupe que le barycentre le plus proche, un nouveau barycentre est calculer pour chaque groupe puis en recalcule le distance de tous les points par rapport à ces nouveaux barycentre , cela est répéter jusqu'à la convergence et que il n'existe plus de point qui change de cluster.



## 5. Association Rule

### 5.1 Définition

Une expression d'implication de la forme  $X \rightarrow Y$  par exemple :

$\{\text{Lait, pain}\} \rightarrow \{\text{fromage}\}$ , cette dernier expression nous révèle que les personnes qui achètent du lait et du pain ensemble vont probablement acheter du fromage aussi, ce genre d'analyse est très utilisé dans le Data Mining, car il révèle les relation entre les différents éléments du data set en les groupant par fréquence d'occurrence mutuel.

Cela veut dire que beaucoup d'applications dans le monde professionnel l'utilise, surtout les départements sales, marketing, dans l'entreprise ainsi que les domaines de la science comme :

---

Market basket analysis, diagnostique Médicale, Protein sequences dans la recherche génétique, CRM pour les business.

L'application que j'ai participé au développement au sein de Djezzy s'appelle " Product Affinity"

Association Rules utilise des algorithmes de ML spéciales de type apprentissage non supervisé dont on peut citer : Apriori, Eclat, FpGrowth, ...

## 5.2 Notions AR:

- **Support** (Soutien). Cela indique à quel point un ensemble d'éléments est populaire, tel que mesuré par la proportion de transactions dans lesquelles un ensemble d'éléments apparaît.

$$Support(X) = \frac{\text{Nbre d'occurrence de } X \text{ dans les transaction}}{\text{Nbre de transaction}}$$

- **Confidence**(Confiance). Cela indique la probabilité que l'article Y soit acheté lors de l'achat de l'article X, exprimé par  $\{X \rightarrow Y\}$ . Ceci est mesuré par la proportion des transactions avec l'article X, dans laquelle l'article Y apparaît également. L'un des inconvénients de la mesure de confiance est qu'elle pourrait mal représenter l'importance d'une association, si l'un des articles dans l'Item set est très populaire.

$$confidance(X, Y) = \frac{Support(X, Y)}{Support(X)}$$

- **Lift**. Cela indique la probabilité que l'article Y soit acheté lorsque l'article X est acheté, tout en contrôlant la popularité de l'article Y.

$$Lift\{X \rightarrow Y\} = \frac{Support\{X,Y\}}{Support(X).Support(Y)}$$

Un « lift » supérieur à 1 traduit une corrélation positive de X et Y, et donc le caractère significatif de l'association.

## 5.3 Algorithms d'Association Rules

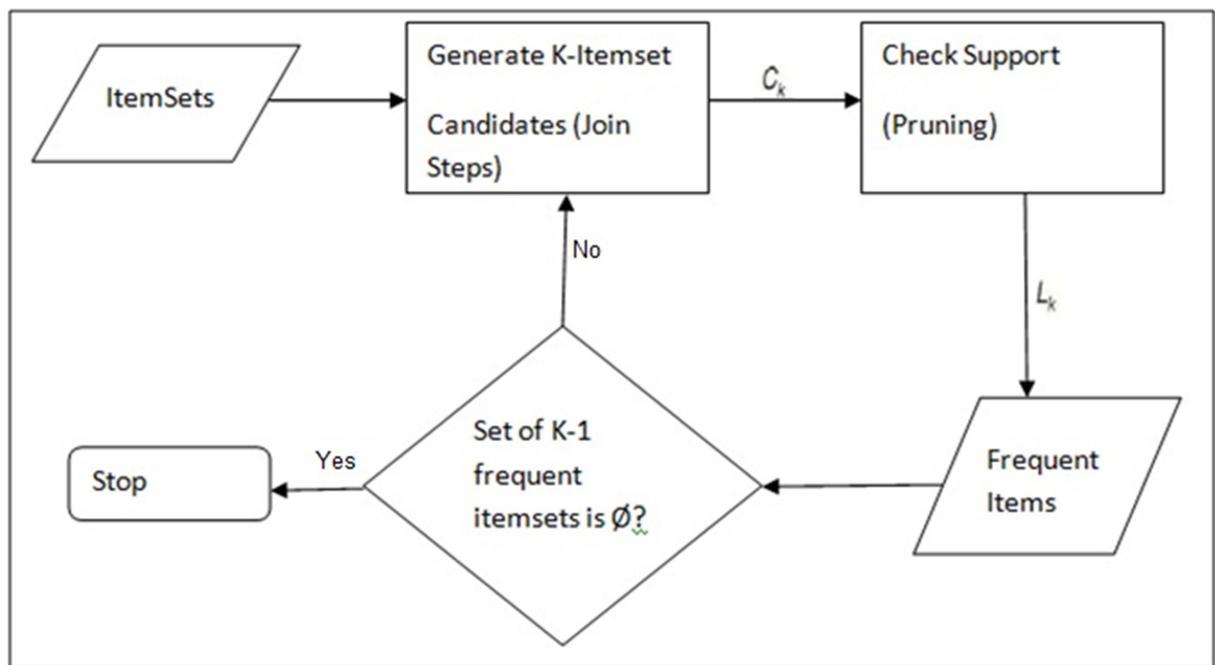
### 5.3.1 Apriori

le principe apriori peut réduire le nombre d'itemsets que nous devons examiner. Dit simplement, le principe apriori stipule que si un ensemble d'items est peu fréquent, alors tous ses sous-ensembles doivent aussi être peu fréquents.

#### Principe :

En utilisant le principe apriori, le nombre d'itemsets à examiner peut être élagué, et la liste des itemsets populaires peut être obtenue dans ces étapes:

- Étape 0. Commencez avec des éléments contenant un seul élément.
- Étape 1. Déterminez le support pour les itemsets. Conservez les ensembles d'éléments qui répondent à votre seuil de support minimal et supprimez les ensembles d'éléments qui ne le sont pas.
- Étape 2. À l'aide des jeux d'éléments que vous avez conservés à l'étape 1, générez toutes les configurations d'ensemble d'éléments possibles.
- Étape 3. Répétez les étapes 1 et 2 jusqu'à ce qu'il n'y ait plus de nouveaux itemsets.



### 5.3.2 ECLAT

ECLAT (equivalent class transformation)

Apriori et FP-growth utilisent un format de données horizontal. Les données peuvent également être représentées en format vertical. Cette Méthode est utilisée par l'algorithme ECLAT.

Eclat: algorithme

1. Obtenir la liste d'articles pour chaque article (BD scan)
2. La liste Tid de {a} est exactement la liste des transactions contenant {a}
3. Intersecté la liste Tid de {a} avec les listes de tous les autres éléments, résultant en des listes Tid de {a, b}, {a, c}, {a, d}, ...  
{a} - base de données conditionnelle (si {a} a été supprimée)
4. Répétez à partir de 1 sur {a} - base de données conditionnelle
5. Répétez pour tous les autres éléments

TID	Items
1	Bread,Butter,Jam
2	Butter,Coke
3	Butter,Milk
4	Bread,Butter,Coke
5	Bread,Milk
6	Butter,Milk
7	Bread,Milk
8	Bread,Butter,Milk,Jam
9	Bread,Butter,Milk

Item Set	TID set
Bread	1,4,5,7,8,9
Butter	1,2,3,4,6,8,9
Milk	3,5,6,7,8,9
Coke	2,4
Jam	1,8

---

### Frequent 1-itemsets

min\_sup=2

Item Set	TID Set
Bread	1,4,5,7,8,9
Butter	1,2,3,4,6,8,9
Milk	3,5,6,7,8,9
Coke	2,4
Jam	1,8

### Frequent 2-itemsets

Item Set	TID set
{Bread,Butter}	1,4,8,9
{Bread,Milk}	5,7,8,9
{Bread,Coke}	4
{Bread,Jam}	1,8
{Butter,Milk}	3,6,8,9
{Butter,Coke}	2,4
{Butter,Jam}	1,8
{Milk,Jam}	8

## Frequent 3-itemsets

Item Set	TID Set
{Bread,Butter,Milk}	8,9
{Bread,Butter,Jam}	1,8

### Les Avantages

- La recherche en première ligne réduit les besoins en mémoire
- Habituellement (considérablement) plus rapide qu'Apriori
- Pas besoin de scanner la base de données pour trouver le support de  $(k + 1)$  itemsets, pour  $k \geq 1$

### Les Inconvénients

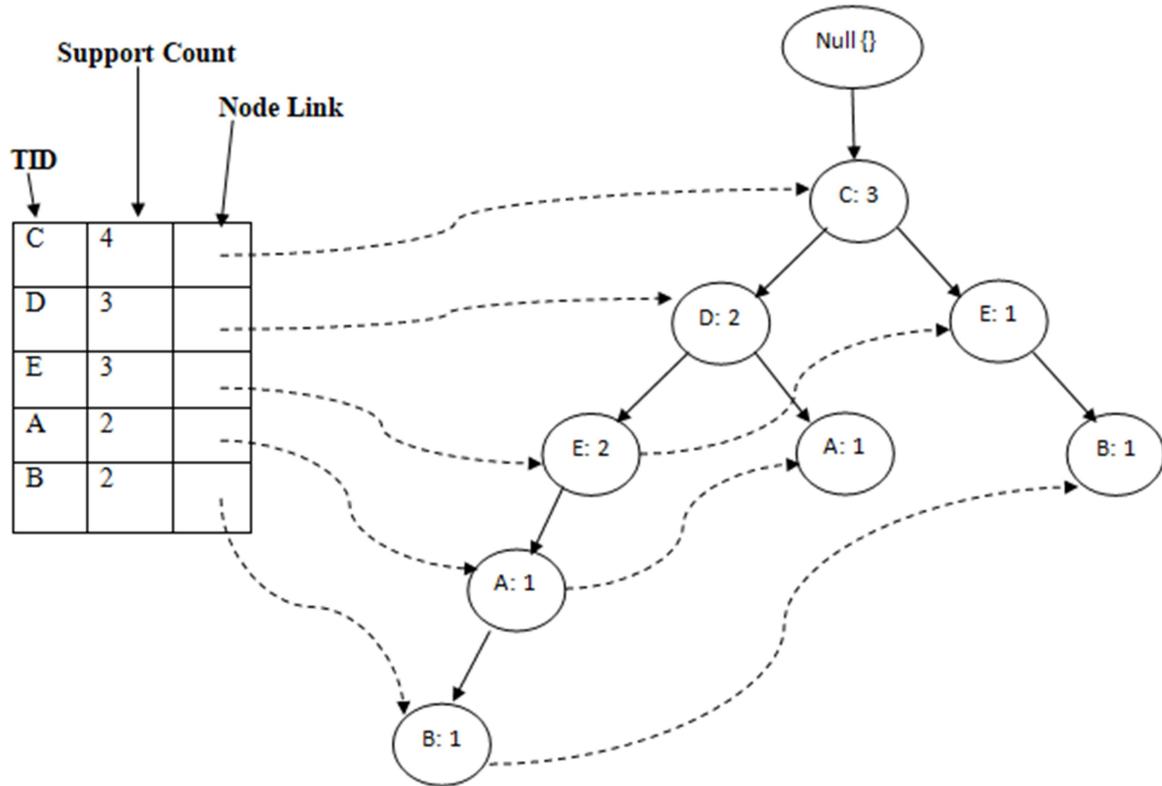
- Les listes TID peuvent être assez longues, donc coûteux à manipuler

### 5.3.3 FpGrowth

L'algorithme Fp-growth permet la découverte des itemsets fréquents sans génération des itemsets candidats.

Le processus se déroule en deux étapes, une étape de construction des arbres FP-tree et une étape d'extraction des item sets fréquents directement de ces arbres. La construction de l'arbre FP-tree s'effectue suivant les étapes ci-dessous :

1. Calculer le support minimal.
2. Calculer chacune des occurrences d'un item constituant la base de transactions.
3. Établir un critère de priorité pour ces items.
4. Faire le tri des items en fonction de leur priorité.
5. Établir le nœud racine.
6. À partir de chaque nœud père insérer les enfants en partant du nœud racine
7. Valider la structure de l'arbre FP-Growth

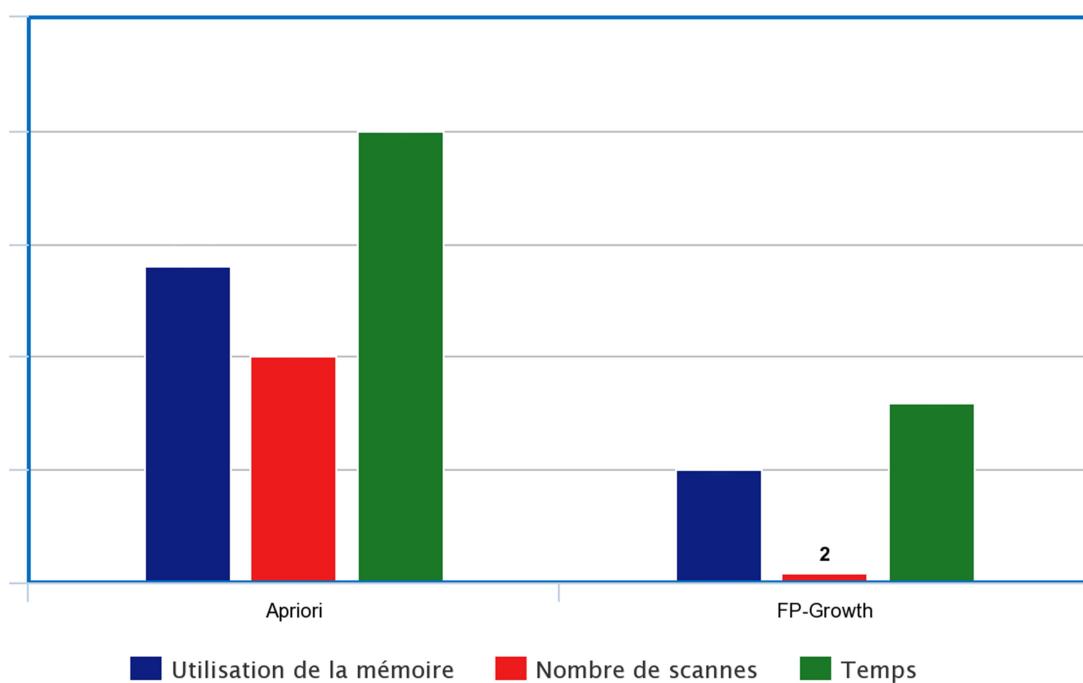


#### 5.4 Comparaison entre l'algorithme APRIORI et algorithme Fp-Growth :

	Apriori	Fp-Growth
Technique	Utilise la propriété d'Apriori et la jointure pour trouver les Itemset Fréquents.	Construit une base de modèle conditionnel (Arbre FP) à partir de la base de données qui satisfait le support minimum ( <b>minsup</b> ).
Type de Stockage	Tableau (Array)	Arbre (Tree)
Utilisation de la mémoire	nécessite un grand espace mémoire car il gère un grand nombre de générations d'ensembles d'éléments candidats	nécessite moins de mémoire en raison de sa structure compacte, ils découvrent le jeu d'éléments fréquents sans génération de jeu d'éléments candidats
Nombre de scannes	effectue plusieurs analyses pour générer un ensemble de candidats (K+1)	analyse la base de données seulement deux fois

<b>Méthode de Recherche</b>	Breadth First search (Traditionnelle)	Divide & Conquer
<b>Temps</b>	le temps d'exécution est plus gaspillé à produire des candidats à chaque fois	inférieur à celui d'Apriori

Apriori Vs. FP-Growth



## 6. Product and Business Affinity

L'affinité produit (Product Affinity) signifie l'appréciation naturelle des clients pour les produits. La segmentation par affinité des produits divise les clients en groupes basés sur les produits achetés.

La compréhension des relations entre les produits, telle que déterminée par les modèles d'achat des clients, peut donner des indications potentiellement utiles sur le comportement des clients et leurs besoins. Une telle information peut être utile pour découvrir des opportunités de ventes croisées et pour déterminer comment promouvoir plus efficacement ses produits. Dans un contexte B2B, où une partie importante des activités d'un organisme de vente peut être orientée vers la fourniture d'informations détaillées sur les produits et le support des applications, la possibilité d'introduire un nouveau produit approprié crée de nouvelles opportunités de vente.

---

Alors que les relations avec les produits peuvent parfois être découvertes par le biais d'observations occasionnelles, de preuves anecdotiques ou d'informations directement fournies par le client, ces approches non systématiques peuvent souvent passer à côté de relations subtiles (mais néanmoins significatives et potentiellement exploitables). Lorsque l'on dispose de bases de données de clients et de leurs achats, ou que l'on peut facilement les assembler, une autre option consiste à utiliser des techniques d'exploration de données pour identifier des relations de produits potentiellement intéressantes. Ce livre blanc décrit comment une telle technique, l'analyse d'affinité, fonctionne et comment elle peut être appliquée pour découvrir des relations exploitables dans les données de produit.

### ***6.1 Le Up-Selling, Down-Selling, Cross-Selling***

La promotion (Up-Selling) consiste à encourager les clients à acheter un produit haut de gamme comparable à celui en question, tandis que la vente croisée (Cross-Selling) invite les clients à acheter des articles connexes ou complémentaires.

En pratique, les grandes entreprises associent généralement l'Up-Selling et le Cross-Selling pour maximiser les profits.

La vente à découvert (Down-Selling) est une autre technique utiliser lorsqu'un client essaie de se retirer d'un achat. En respectant le budget du client et lui donnant un meilleur prix (moins cher qu'avant suggéré) pour un produit / service qui a des caractéristiques familières comme l'autre avant cela. Dans ce cas, on offre un produit moins cher, qui a plus de chances d'être accepté, car vendre quelque chose vaut mieux que rien.

### ***7. Le cas d'utilisation du de Product affinity au sein de Djezzy***

Djezzy possède plusieurs plateformes de gestion de l'information, un entrepôt de données (Data warehouse Teradata)) qui contient plus de 6000 tables, d'un volume dépassant 47 TB plus une plateforme Big Data (Horton Works) se composant de plus de 45 serveurs.

Notre travail se focalise sur l'étude de différentes offres proposé par l'entreprise, les achats de ces offres peuvent être trouvés sur la plateforme, avec une moyenne de 16M de transactions par trimestre pour 8M d'abonnées.

Utilisant l'enivrement « Hadoop Hive + Apache Spark », Depuis cette Platform on a extrait un data Set de 16 Millions de transactions touchant 8,300,000 clients pour les dernier 3 mois .On fait une analyse de ces données utilisant « Association Rules FP-Growth Algorithm ».On utilisant la bibliothèque « Orange3-Associate » Pour l'implémenter sur python3 .

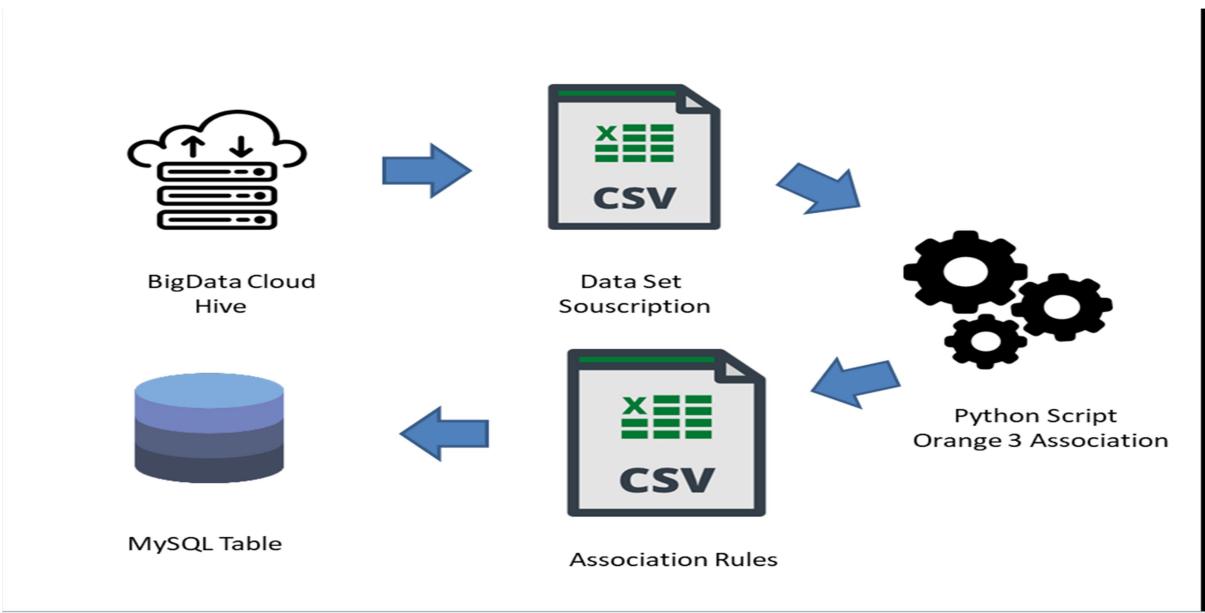
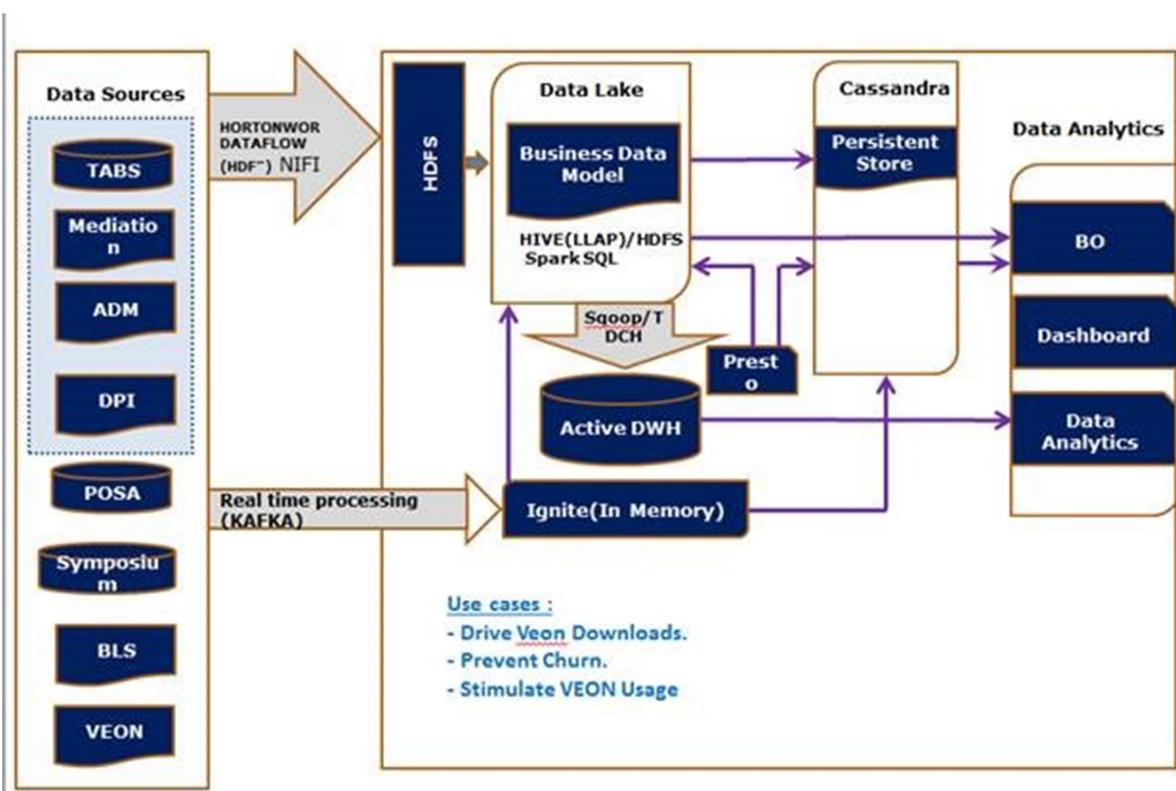


Figure 5 flow of work

```

1 # -*- coding: utf-8 -*-
2 """
3 Created on Sun Jul 29 11:42:44 2018
4
5 @author: KerMA
6 """
7 from orangecontrib.associate.fpgrowth import *
8 import csv
9
10 ##### Loading Data From Data Set #####
11
12 itemset=[]
13 fhand = open('itemsets.csv')
14 for line in fhand:
15     line =line.split(" ")
16     for x in range(0,len(line)):
17         line[x] = line[x].strip('\n')
18     itemset.append(line)
19
20 ##### Generating Association Rules #####
21 ##### using FP-Growth Algorithm #####
22
23 a=[]
24 itemsets = dict(frequent_itemsets(itemset, 50000))
25 rules = association_rules(itemsets, .7)
26 N=8286765
27 for x in rules_stats(rules, itemsets, N):
28     a.append(list(x))
29
30 ##### Writing Rules to CSV #####
31
32
33 with open('prod_aff_bigexp.csv', 'w') as csvfile:
34     writer = csv.writer(csvfile, delimiter=',', quoting=csv.QUOTE_MINIMAL)
35     for i in a:
36         writer.writerow(i)
37

```

On A Sélectionner Tous Les règles Avec un support supérieur A 50000 client « Le Support est D'habitude Exprime en pourcentage mais la Bibliothèque précise l'utilisation du nombre d'instances » « 0.6% », et une Confidence supérieure 70%. Il faut noter que Djazzy à Plus de 16 Millions d'abonnés.

Le Résultat est enregistré sur un fichier CSV, Puis Exporter Sur Une Base de Données MySQL,

Dans les résultats on constat que on ne trouve pas une règle avec un Lift supérieur à 1, Un Lift > 1 Veut dire que l'item set antécédent a un effet positif sur le item set conséquent. Alors on ne trouve pas une règle exploitable qui représente une opportunité pour le business.

```
SELECT `id`, `antecedent`, `consequent`, `support`, `confidence`, `lift` FROM `prod_aff_sup50000_conf070`
```

<b>id</b>	<b>atecedent</b>	<b>consequent</b>	<b>support</b>	<b>confidence</b>	<b>lift</b>
1	'82', '1', '25'	10	107046	0.817888	0.00166055
2	'82', '1'	10	399647	0.756662	0.00153624
3	'1', '25'	10	195968	0.7298	0.0014817
4	'1', '20'	10	60219	0.705215	0.00143179
5	'82', '22'	25	59604	0.83757	0.00260558
6	'22', '10'	25	71379	0.833137	0.00259179
7	'22', '28'	25	54746	0.721652	0.00224497
8	'22', '20'	25	55592	0.789379	0.00245566
9	'24', '28'	20	68473	0.836628	0.00773353
10	'24', '25'	20	93190	0.802518	0.00741822

On trouve aussi que les produits les plus populaires sont les produits avec un prix inférieur à 100DA.

### **8. Conclusion :**

L'analyse d'**Association Rules** et du **Product Affinity** est un outil utile voire même important pour étudier les règles d'association entre les différents produits et offres. Pour identifier la présence d'opportunités d'accroissement et d'investissement. Et faire une Analyse SWOT au sein de l'entreprise. (Particulièrement utile pour détecter les opportunités et les forces de l'entreprise « S, O »).

---

## Bibliographie

- 1.Data Mining: Concepts and Techniques - 3rd Edition Jiawei Han
- 2.Apriori Algorithm Jung Hoon Kim KAIST Knowledge Service Engineering Data Mining Lab.
- 3.AndrewNg Machine Learning Course (Coursera)
- 4.Machine Learning A-Z - Hands On Python and R In Data Science (Udemy)
- 5.Data Mining: Methodology and Algorithms Ryan A. Rossi Purdue University
- 6.Fall 2004, CIS, Temple University CIS527: Data Warehousing, Filtering, and Mining
- 7.Frequent Pattern (FP) growth Algorithm for Association Rule Mining
- 8.<https://itsocial.fr/articles-decideurs/4-types-de-donnees-internet-des-objets-pour-3-modeles-big-data/>
9. [http://scikit-learn.org/stable/auto\\_examples/svm/plot\\_svm\\_nonlinear.html](http://scikit-learn.org/stable/auto_examples/svm/plot_svm_nonlinear.html)
- 10.<http://scikit-learn.org/stable/modules/clustering.html#k-means>
- 11.(Eclat Association Rule Learning) <https://www.youtube.com/watch?v=oBiq8cMkTCU>
- 12.Apriori Algorithm (Associated Learning) [https://www.youtube.com/watch?v=WGIMIS\\_Ydkg](https://www.youtube.com/watch?v=WGIMIS_Ydkg)
- 13.<https://www.youtube.com/watch?v=yCbankIouUU&t=15s>
- 14.<https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html>
- 15.[http://orange3-associate.readthedocs.io/en/latest/scripting.html#fpgrowth.rules\\_stats](http://orange3-associate.readthedocs.io/en/latest/scripting.html#fpgrowth.rules_stats)