Université Claude Bernard Lyon 1

- Master 2 Data Science -

Data Mining

Rapport du projet Data Mining
- Etude d'une campagne de marketing -

Binôme :

- **LALMAS Sonia (p2311211)**
- **TSOUALLA TATIDOUNG Grace (p2212516)**

Les membres du binôme ont défini ensemble toutes les tâches et analyses effectuées dans l'étude qui va suivre et le travail sur toutes les parties a été fait en totale collaboration de façon conjointe.

2023/2024

1. Description des données :

L'ensemble de données provient du livre Business Analytics Using SAS Enterprise Guide and SAS Enterprise Miner d'O.Par-Rud.

Il est lié à un contexte marketing, avec pour objectif d'améliorer l'efficacité des campagnes marketing en effectuant des analyses dessus.

Voici une description de quelques variables :

- Dt Customer :Date de l'inscription du client auprès de l'entreprise.
- Education :Niveau d'éducation du client.
- Marital :Statut matrimonial du client.
- Kidhome :Nombre de petits enfants dans le ménage du client.
- Teenhome :Nombre d'adolescents dans le ménage du client.
- Income :Revenu annuel du ménage du client.
- MntFishProducts, MntMeatProducts, MntFruits, MntSweetProducts, MntWines, MntGoldProds : Montant dépensé dans différentes catégories de produits
- AcceptedCmp1..5 : le client a accepté l'offre lors des 1re à 5e campagnes.
- Complain :Variable binaire indiquant si le client s'est plaint au cours des 2 dernières années
- Response (cible) :Variable binaire indiquant si le client a accepté l'offre lors de la dernière campagne.

2. Quelques analyses statistiques :

Nous avons présenté la répartition des clients en fonction de leur niveau d'éducation et de leur statut marital afin d'Identifier les niveaux d'éducation et les catégories de statut marital les plus fréquentes.

Nous pouvons noter que la majorité des clients sont diplômés ou ont obtenu un doctorat avec une majorité de personnes mariés ayant un revenu stable et des familles à nourrir. Cela nous oriente sur le type de client qui vont être traités dans cette étude et nous permet d'affiner nos hypothèses. (voir Annexe 1)

Nous avons aussi effectué une analyse sur la fréquence d'apparition des campagnes de marketing, cette analyse servira à compléter notre étude des patterns fréquents que nous allons effectuer par la suite (voir Annexe 4)

On a analysé les metrics les plus importantes et fournit des statistiques descriptives sur les métriques clés: la moyenne, l'écart-type, la médiane, les quartiles, etc.pour mieux nous familiariser avec les variables et les individus.(voir Annexe 2)

Une analyse des correlation a permi de trouver quelques relations entre certaines variables tel que:

- Le revenu et le montant des dépenses des individus pour plusieurs produits différents
- Le revenu et l'acceptation de la campagne 3
- Le revenu et le nombre d'enfant en bas âge

Cela nous permet de dire que le revenu des individu influe souvent sur leur mode de vie, si nous devant mettre en place des campagne marketing dans le future c'est un critère qu'il est important de prendre en compte.(voir Annexe 3)

3. Traitement des outliers :

Nous avons utilisé l'algorithme "LocalOutlierFactor" afin de détecter les anomalies dans notre dataset, nous avons ensuite déterminer si ces anomalies sont des valeurs aberrantes ou des erreurs (voir Annexe 5)

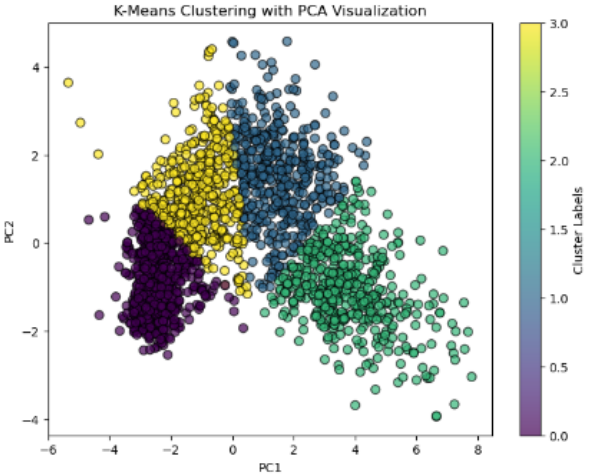
D'après notre analyse, on peut en déduire qu'on a un ensemble de 8 personnes qui ont un revenu annuel de plus de 100000 $ mais qui consomment très peu, sachant que les montants représentent une consommation sur 2 ans. On les considère donc comme des erreurs et nous décidons de les supprimer pour ne pas biaiser notre analyse.(voir Annexe 4)

4. Clustering :

Pour effectuer le processus de clustering il est nécessaire de garder uniquement des données numériques, car l'algorithme k-means utilisé se base sur la distance euclidienne pour mesurer la similarité entre les points.

La méthode du coude nous a permis de se rapprocher le plus possible du nombre de clusters appropriés, dans notre cas elle a donné 4 clusters. (voir Annexe 6)

Afin de renforcer la robustesse de notre analyse, nous avons commencé par effectuer une normalisation et une analyse en composantes principales (PCA) avant d'appliquer l'algorithme K-means.



Cette étude a permis de classer les client en 4 catégories :

- le catégorie 1 :

Ces client ont un revenu élevé et et leurs dépenses sont tout autant élevés
Il ne possède pas énormément d'enfants en bas âge ni d'adolescent
La campagne à laquelle ils participent le plus est la campagne 3

- le catégorie 2 :

Ces client ont un revenu moins élevé que la précédente et et leurs dépenses sont en accord avec leur revenu
Ils possèdent un grand nombre d'adolescents
Ils ne semblent intéressés par aucune des compagne proposée en particulier

- le catégorie 3 :

Ces client ont un revenu plutôt faible
Ils possèdent un grand nombre d'adolescents
Ils ne semblent intéressés par aucune des compagne proposée en particulier

- le catégorie 4:

Ces client ont un revenu très faible
Ils possèdent énormément d'enfants en bas âge
Ils ne semblent intéressés par aucune des compagne proposée en particulier

Cette classification nous permet de mieux orienter les prochaines campagnes marketing et d'en créer de nouvelles plus adaptées à la clientèle de ce supermarché, étant donné qu'on remarque un intérêt faible de la part des 3 dernières catégories pour les compagnes qui sont proposées actuellement.

5. détection de communautés :

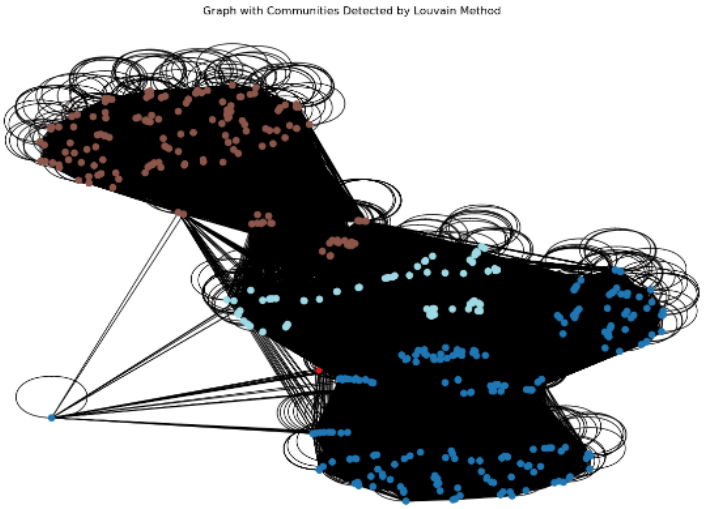
- Création du graphe

On souhaiterait également tester une approche de clustering en utilisant les graphes car nous considérons que les données sur les campagnes peuvent permettre de détecter des communautés intéressantes.

Ainsi on crée un graphe tel qu' un nœud est une personne du jeu de donnée identifié par son ID et il y a un lien entre les personnes ayant accepté les mêmes campagnes. Le lien étant pondéré par le nombre de campagnes acceptées. (voir Annexe 7)

L'application de l'algorithme de Louvain met en avant 3 communautés qu'il est intéressant d'exploiter lorsqu'on souhaite lancer une campagne de marketing.

Ces communautés nous aident à entrevoir les clients qui ont les mêmes intérêts pour les communautés et d'adapter les nouvelles compagne selon leurs profil.



- **Un petit monde :**

Sur le graphe obtenu nous avons effectué quelques statistiques qui ont révélé que le graphe obtenu est un “petit monde” , on peut en déduire que l'idéal pour ce supermarché en terme de stratégie marketing est de cibler les noeuds/personnes les plus influentes du réseau pour sa campagne marketing et elle sûre qu'elle pourra atteindre sans problèmes les autres clients.

- **Analyses des personne non connectées au reste :**

On remarque que le graphe contient uniquement 459 personnes, ce qui signifie que 1749 personnes sont manquantes. On n'émet alors l'hypothèse selon laquelle, ce sont des personnes qui n'ont accepté aucune campagne et dans une autre partie, nous essayerons de comprendre pourquoi, ce qui les caractérisent, ce qu'ils ont en commun dans une autre partie. (voir Annexe 8)

Une analyse sur ces individus permet d’affirmer que ce sont des personnes qui dans l'ensemble font très très peu d'achats surtout que c'est une consommation sur deux ans, on peut donc dire que ce n'est certainement pas leur supermarché principal ou bien qu'ils n'y retrouvent pas tous les produits recherchés.

On peut donc émettre une hypothèse selon laquelle ce supermarché a du mal à fidéliser ou qu'elle ne cible pas les bons clients lors de ses campagnes car ceux qui n'ont accepté aucune campagne représentent presque 80% du jeu de données. Elles devraient donc réadapter les campagnes en fonction du profil de ses clients.

6. Frequent patterns :

Nous avons effectué une analyse des patterns fréquents qui peuvent apparaître par rapport aux campagnes de marketing effectuées. (voir Annexe 9)

- Ce que nous pouvons affirmer d'après les résultats, les compagnes 4, 3 et 5 sont les plus fréquentes.
- Nous pouvons aussi dire que les clients qui acceptent la campagne 5 ont tendance à accepter la campagne 1 et vise versa.

De cette étude il sera possible de déterminer les campagnes qui ressortent le plus souvent et qui sont le plus acceptées par les clients du supermarché, celles-ci vont être gardées, pour celles qui ne sont pas souvent acceptées il serait préférable de les supprimer ou de modifier pour mieux satisfaire les clients.

7. Le système de recommandation:

Nous souhaitons utiliser les approches de systèmes de recommandations afin de prédire la colonne response qui vaut 1 si la personne a accepté la dernière campagne et 0 sinon. On crée un nouveau jeu de données avec les campagnes et la colonne Response. (voir Annexe 10)

On utilise deux approches :

- **USER_BASED KNN :**

Afin de prédire la colonne réponse pour une personne donnée, on va chercher ses plus proches voisins(ceux qui lui sont le plus similaire) en calculant un critère de similarité entre lui et toutes les autres personnes. Nous avons choisi la similarité cosinus comme critère dans notre cas.

Une fois les plus proches voisins trouvés, on va par la suite faire la moyenne des votes de ces derniers pondérés par la valeur du critère calculé afin de trouver celui de la personne qui nous intéresse.

- **NMF - NON NEGATIVE MATRIX FACTORISATION :**

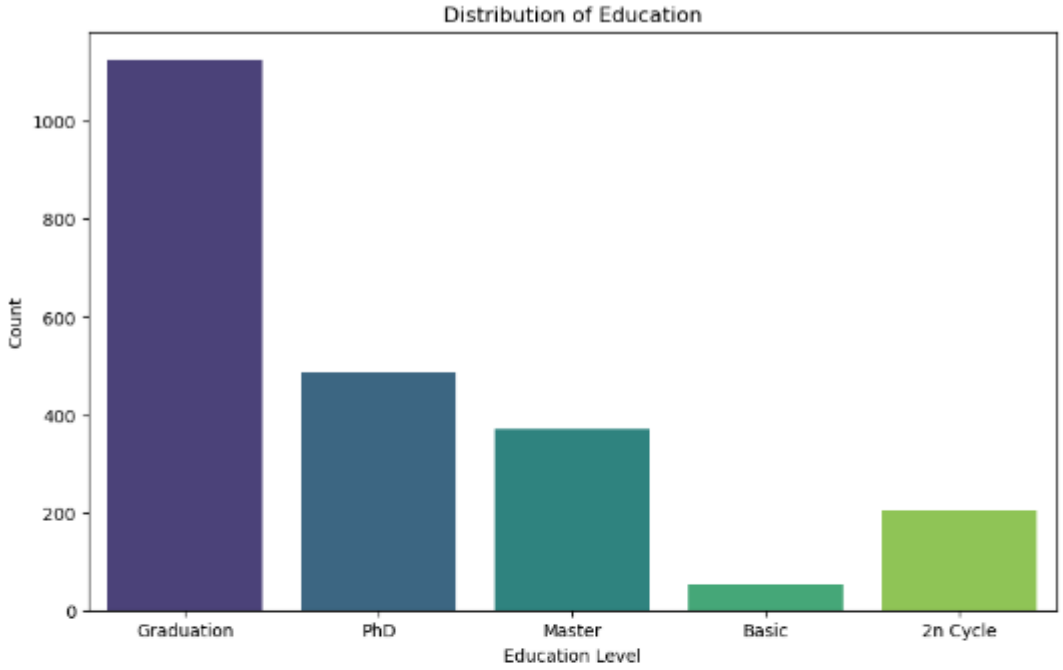
On va essayer de chercher des variables latentes qui caractérisent notre jeu de données. Comme dans notre cas, on n'a pas d'informations sur le contenu des campagnes. On va supposer que chaque campagne est destinée à mettre en avant un produit.

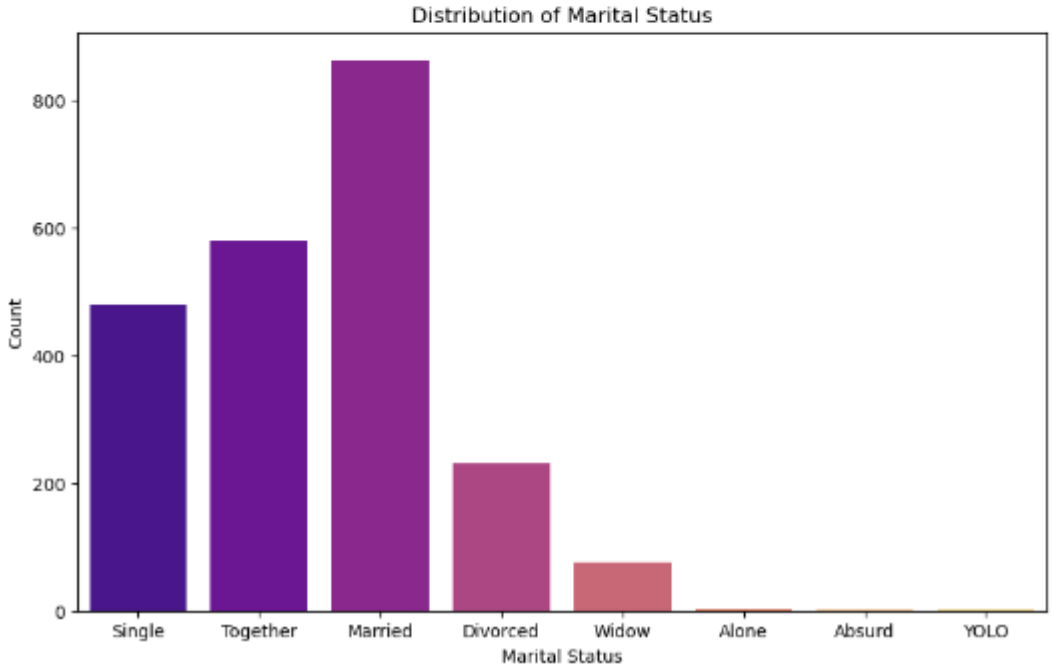
On va donc chercher 5 variables latentes qui correspondent à nos 5 produits et chercher deux matrices W qui vont caractériser nos utilisateurs par rapport à nos variables latentes et une matrice H qui va caractériser nos campagnes par rapport à nos variables latentes. On obtient donc des vecteurs pour chaque utilisateur et chaque campagne. On va multiplier ces vecteurs par utilisateur et pas campagne. On recommande donc selon la plus grande valeur obtenue dans la ligne pour une personne donnée.

- On remarque aussi que la propriété de petit monde de notre graphe est particulièrement bénéfique pour notre système de recommandation car les performances obtenues sont très encourageantes.

Annexes :

Annexe 1 :



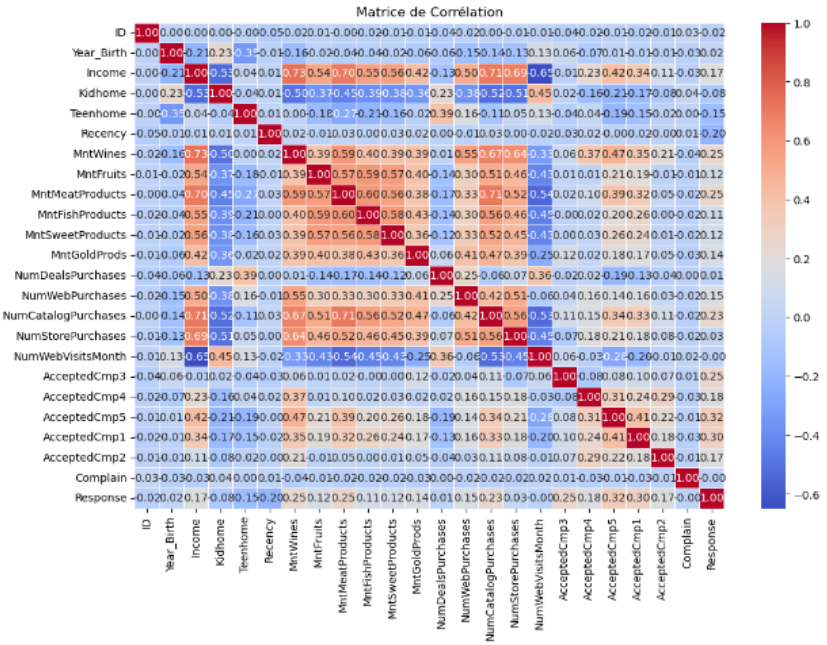


Annexe 2 :

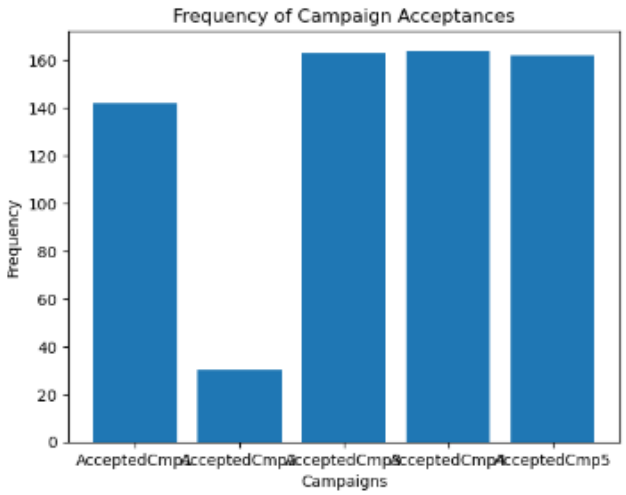
	ID	Year_Birth	Income	Kidhome	Teenhome	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	...	Nu
count	2240.000000	2240.000000	2216.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	2240.000000	...	
mean	5592.159821	1968.805804	52247.251354	0.444196	0.506250	49.109375	303.935714	26.302232	166.950000	37.525446	...	
std	3246.662198	11.984069	25173.076661	0.538398	0.544538	28.962453	336.597393	39.773434	225.715373	54.628979	...	
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	
25%	2828.250000	1959.000000	35303.000000	0.000000	0.000000	24.000000	23.750000	1.000000	16.000000	3.000000	...	
50%	5458.500000	1970.000000	51381.500000	0.000000	0.000000	49.000000	173.500000	8.000000	67.000000	12.000000	...	
75%	8427.750000	1977.000000	68522.000000	1.000000	1.000000	74.000000	504.250000	33.000000	232.000000	50.000000	...	
max	11191.000000	1996.000000	666666.000000	2.000000	2.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	...	

8 rows × 26 columns

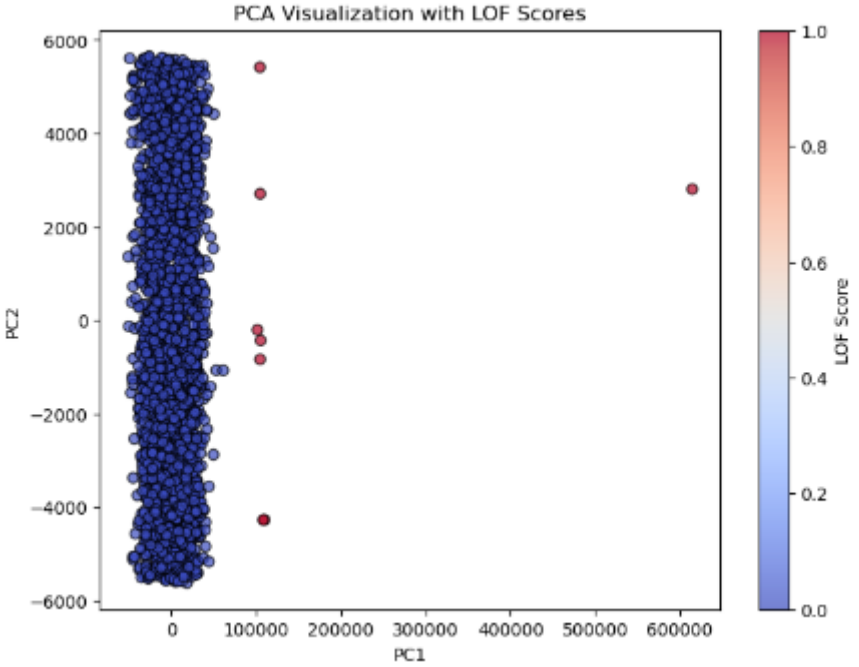
Annexe 3 :



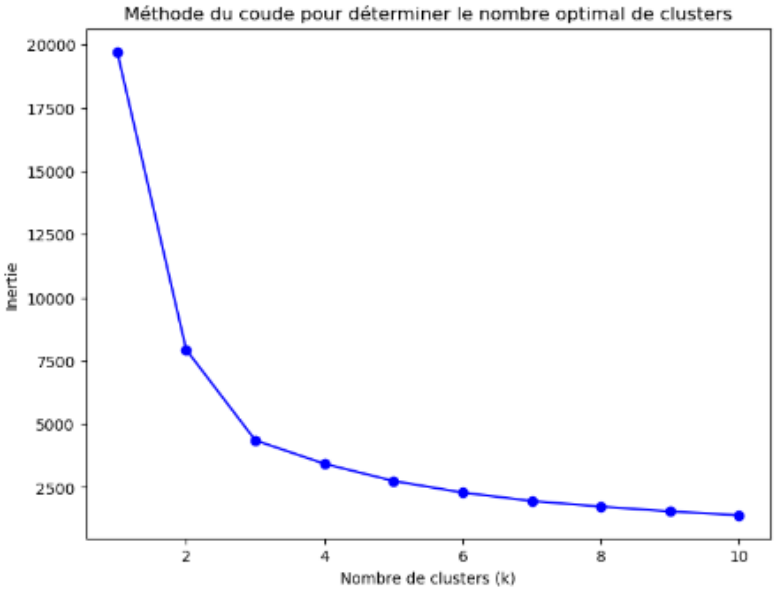
Annexe 4 :



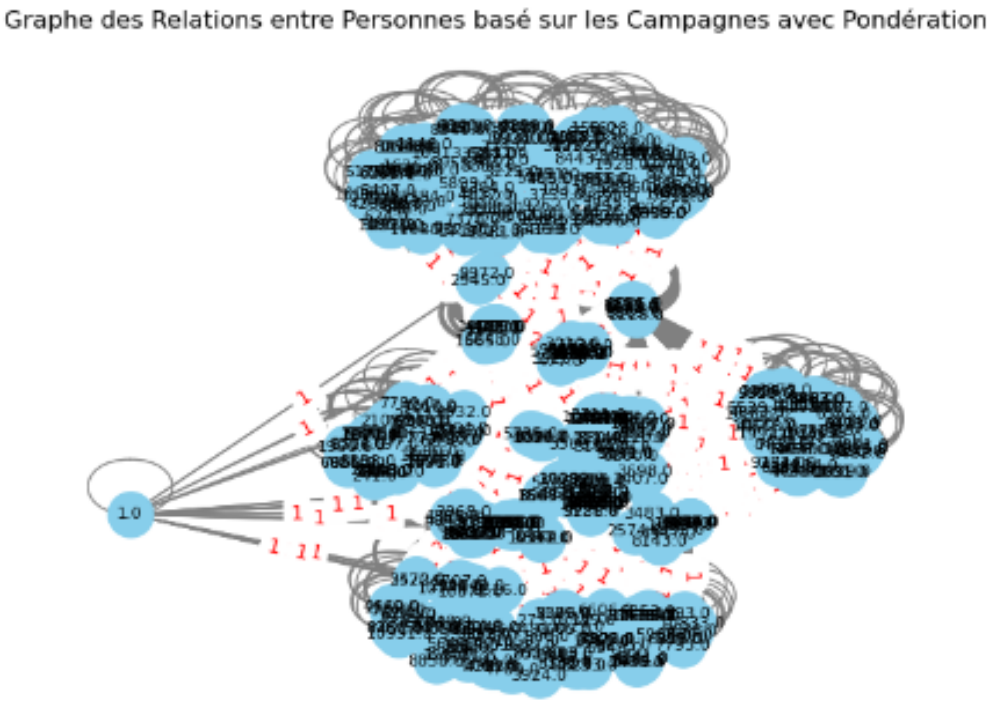
Annexe 5 :



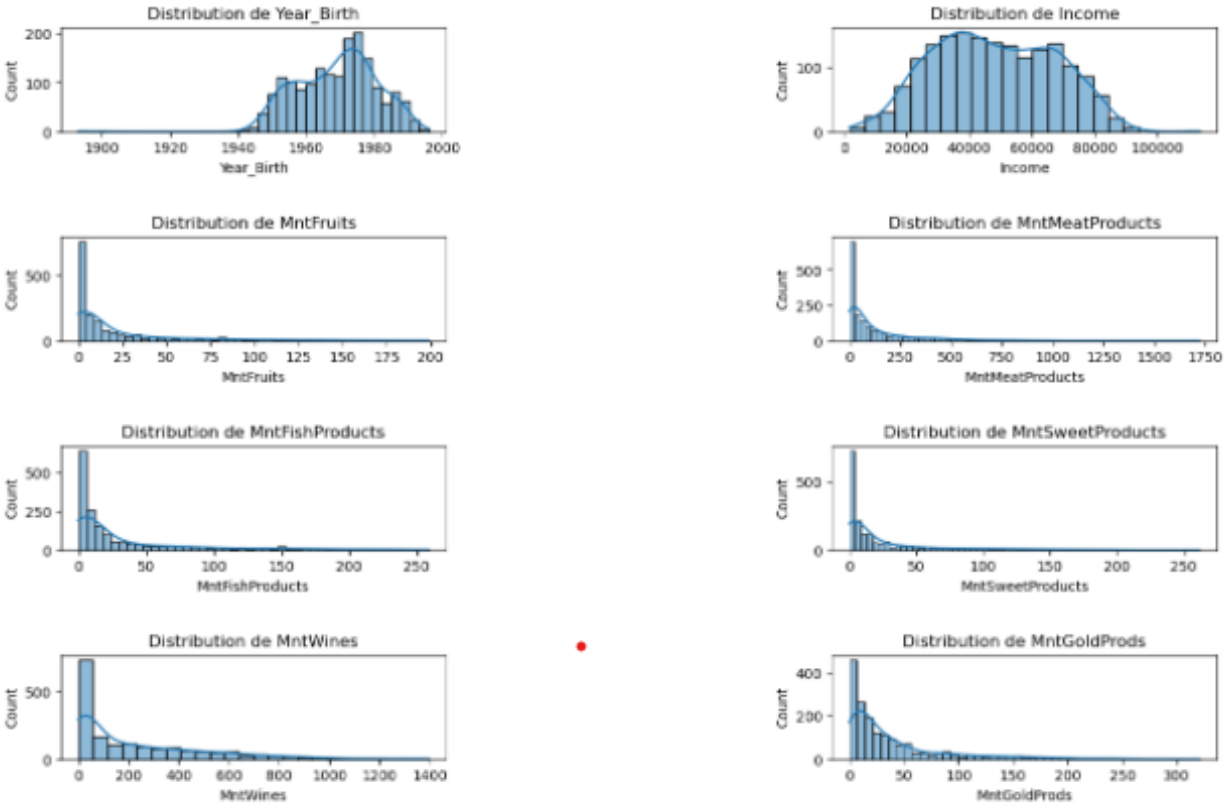
Annexe 6 :



Annexe 7 :



Annexe 8 :



Annexe 9:

	support	itemsets
0	0.073822	(AcceptedCmp3)
1	0.074275	(AcceptedCmp4)
2	0.073370	(AcceptedCmp5)
3	0.064312	(AcceptedCmp1)
4	0.013587	(AcceptedCmp2)
5	0.010870	(AcceptedCmp5, AcceptedCmp3)
6	0.010870	(AcceptedCmp3, AcceptedCmp1)
7	0.026721	(AcceptedCmp5, AcceptedCmp4)
8	0.020380	(AcceptedCmp1, AcceptedCmp4)
9	0.030797	(AcceptedCmp5, AcceptedCmp1)
10	0.014040	(AcceptedCmp5, AcceptedCmp1, AcceptedCmp4)

	antecedents	consequents
0	(AcceptedCmp5)	(AcceptedCmp3)
1	(AcceptedCmp3)	(AcceptedCmp5)
2	(AcceptedCmp3)	(AcceptedCmp1)
3	(AcceptedCmp1)	(AcceptedCmp3)
4	(AcceptedCmp5)	(AcceptedCmp4)
5	(AcceptedCmp4)	(AcceptedCmp5)
6	(AcceptedCmp1)	(AcceptedCmp4)
7	(AcceptedCmp4)	(AcceptedCmp1)
8	(AcceptedCmp5)	(AcceptedCmp1)
9	(AcceptedCmp1)	(AcceptedCmp5)
10	(AcceptedCmp5, AcceptedCmp1)	(AcceptedCmp4)
11	(AcceptedCmp5, AcceptedCmp4)	(AcceptedCmp1)
12	(AcceptedCmp4, AcceptedCmp1)	(AcceptedCmp5)
13	(AcceptedCmp5)	(AcceptedCmp4, AcceptedCmp1)
14	(AcceptedCmp1)	(AcceptedCmp5, AcceptedCmp4)
15	(AcceptedCmp4)	(AcceptedCmp5, AcceptedCmp1)

Annexe 10 :

```
Précision du modèle : 80.32%
Matrice de confusion :
[[319  56]
 [ 31  36]]
Rapport de classification :
              precision    recall  f1-score   support

           0       0.91      0.85      0.88       375
           1       0.39      0.54      0.45        67

    accuracy                           0.80       442
   macro avg       0.65      0.69      0.67       442
weighted avg       0.83      0.80      0.82       442
```