Cours d'introduction à l'analyse de données

Mini-projet : Segmentation client sur une base de données de campagne marketing

I. Un peu de contexte

L'analyse de la personnalité du client est une analyse détaillée des clients idéaux d'une entreprise. Elle permet à une entreprise de mieux comprendre ses clients et de modifier plus facilement ses produits en fonction des besoins, des comportements et des préoccupations spécifiques des différents types de clients. Elle permet un meilleur ciblage client pour peut analyser quel segment de clients est le plus susceptible d'acheter le produit, puis commercialiser le produit uniquement auprès de ce segment particulier.

II. Présentation de la base de données

Voici un descriptif des attributs fournis pour les données de la base de données marketing campaign.

Population

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- Income: Customer's yearly household income
- Kidhome: Number of children in customer's household
- Teenhome: Number of teenagers in customer's household
- Dt_Customer: Date of customer's enrollment with the company
- Recency: Number of days since customer's last purchase
- Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company's website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company's website in the last month

➔ **<u>Objectif :</u> Proposer un clustering pertinent permettant de cerner la segmentation client pour cette entreprise.**

III. <u>Analyse de la base de données</u>

NB : ce sont des questions indicatives et vous êtes aussi libres d'analyser la base autrement.

1. Effectuer une phase de nettoyage de la base de données (typage des dates, valeurs manquantes, valeurs aberrantes, renommer les variables par souci de clarté)

2. Créer des variables *age, spent, living with, children, family_size ,is_parent, education2, customer_for* et éliminer les redondances occasionnées.

- *spent : montal total dépensé par le client dans les diverses catégories sur 2 ans*
- *living with : statut simplifié en couple / célibataire*
- *children : nombre total d'enfants du foyer*
- *family_size : nombre de total d'habitants du foyer*
- *is_parent : 0 si non, 1 si oui*
- *education2 : niveau d'éducation simplifié undergrad/grad/postgrad*
- *customer_for : nombre de jours depuis que le client a commencé ses achats au magasin depuis la dernière date enregistrée*
- *age : âge actuel du client*

3. Réaliser une analyse descriptive et exploratoire pertinente en vous concentrant sur certaines variables (présenter les graphiques ou métriques adéquates et interprétez-les).

4. Phase de preprocessing :

   i. Quelles sont les variables catégorielles ? A l'aide de la fonction Python LabelEncoder() s'assurer de les ré-encoder de manière numérique.
   ii. Effectuer un *scaling* en utilisant la fonction *StandardScaler()* sur une copie de la base de données où l'on aura pris soin d'enlever les caractéristiques sur les promos.
   iii. Justifier l'intérêt de ces pratiques courantes en data analyse.

5. Réduction de la dimensionnalité et clustering :

Dans ce problème, il existe de nombreux facteurs sur la base desquels la classification finale sera effectuée. Ces facteurs sont essentiellement des attributs ou des caractéristiques. Plus le nombre de caractéristiques est élevé, plus l'analyse est complexe. Nombre de ces caractéristiques sont corrélées si ce n'est redondantes, et l'on souhaite réduire la dimensionnalité du problème avant d'utiliser des méthodes de classification.

*La **réduction de la dimensionnalité** est le processus qui consiste à réduire le nombre de variables aléatoires considérées, en obtenant un ensemble de variables principales. Une méthode classique de réduction de la dimensionnalité est la méthode **Principal Component Analysis (PCA)**, permettant d'améliorer l'interprétation tout en minimisant la perte d'informations.*

Après avoir réduit la dimension, on peut chercher à effectuer des regroupements de nos données. Différentes méthodes existent, comme la méthode des K-means qui minimise les distances entre les membres d'un cluster avec un nombre initial pré-défini de clusters ou encore la méthode Agglomerative Clustering qui initialement propose chaque point comme un cluster séparé avant de regrouper les pairs de clusters les plus proches jusqu'à arriver à un nombre donné de clusters.

i. A l'aide de la fonction *PCA()* faite une analyse de la base réduite en considérant seulement 3 dimensions. Faire apparaître une figure de la projection des données sur ces 3 dimensions.
ii. Tester et choisir une des deux méthodes de clustering avec différents paramètres de nombres de clusters.

⇨ **from sklearn.cluster import KMeans**
⇨ **from sklearn.cluster import AgglomerativeClustering**

Faire fitter le modèle de clustering choisi sur la PCA. Visualiser par un 3D-plot.

6. Évaluation des modèles et interprétation du profil consommateur

i. Les clusters divisent-ils les clients de manière régulière (ex : avec 4 clusters, a-t-on à peu près 25% des clients dans chaque cluster) ?
ii. Comment sont regroupés les clients par caractéristiques croisées pertinentes ? (ex : dépense/revenu)
iii. Étudier la répartition des clusters sur les différentes campagnes passées.
iv. Étudier les déterminants de la dépense en fonctions des variables disponibles et en déduire les profils consommateurs définis par les clusters.