TP EPITA:

KANTAR

Arnaud Baradat - Quentin Fisch -
Théo Ripoll - Bastien Pouëssel

https://github.com/QuentinFISCH/qantar

Sommaire

01 **Clusterisation**

Méthodes de clusterisation choisies

02 **Personas**

Présentation des clusters à travers des personas

03 **Golden questions**

Réaffectation en fonction du nombre de questions posées

04 **Variables illustratives**

Réaffectation en inversant usage et attitude entre questions et clustering

01

Clusterisation

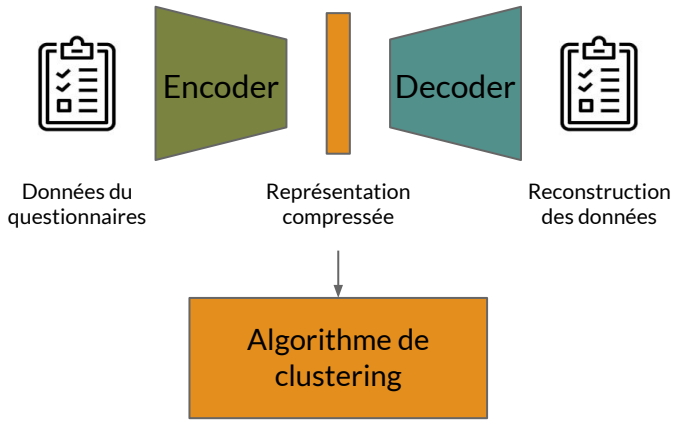
Méthodes de clusterisation choisies

PCA vs Deep Clustering

En ce qui concerne le prétraitement et la représentation des données que nous allons utiliser dans notre clustering, nous avons employé trois approches :

- Les données brutes
- L'utilisation de l'analyse en composantes principales
- L'utilisation d'un auto-encodeur pour la compression des données

Approches deep clustering:



L'utilisation de la PCA a surpassé les données brutes avec un meilleur score de silhouette. Le Deep Clustering, utilisant un auto-encodeur pour compresser les données, vise à extraire des caractéristiques complexes.

L'approche PCA affiche un **score d'environ +3** supérieur au Deep Clustering, qui est moins efficace en raison d'un volume de données insuffisant pour une approche d'apprentissage profond. Nous utilisons donc l'approche PCA.

Preprocessing

OneHotEncoder	Transformation des données en colonnes
Poids	Application du poids présent dans la data
Standardisation	Standardisation des données numériques
PCA	PCA à 3 dimensions pour la visualisation

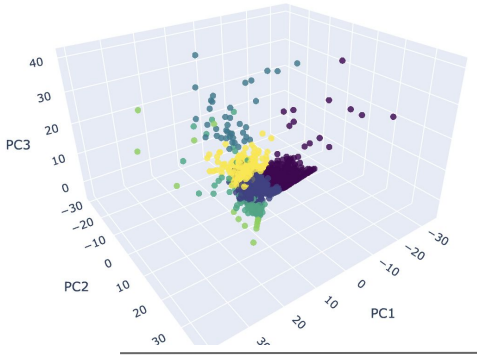
Méthodes de clusterisation testées

Il nous a été demandé de créer entre 4 à 10 groupes de clients ayant répondus au questionnaire. En utilisant la méthode du coude, nous trouvions initialement un nombre de clusters optimal à 3. Nous avions finalement choisi de générer 6 clusters (afin de suivre les consignes) à travers plusieurs méthodes de clusterisation:

1. BIRCH
2. DBSCAN
    a. Très mauvais résultats, ne seront pas présentés dans ces slides
3. Agglomerative Clustering
4. KMeans
5. Autoencoder avec espace latent:
    - The **fonts and colors** used in the template.
    - A selection of **illustrations**. You can also customize and animate them as you wish with the online editor. Visit **Storyset** to find more.
    - More **infographic resources**, whose size and color can be edited.
    - Sets of **customizable icons** of the following themes: general, business, avatar, creative process, education, help & support, medical, nature, performing arts, SEO & marketing, and teamwork.

L'évaluation des clusters a été faite grâce au score de silhouette, la répartition uniforme des clusters ainsi qu'une analyse qualitative

Attitude des consommateurs



Birch

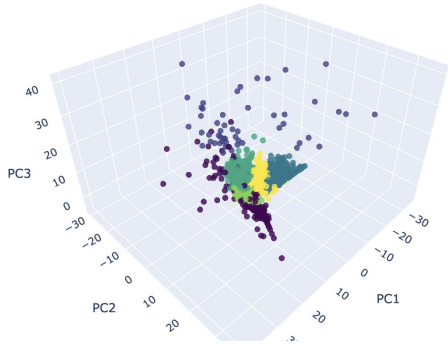
- Silhouette: 0.213
- Répartition: 707, 3096, 39, 125, 15, 1018
- Qualité: Bonne séparation

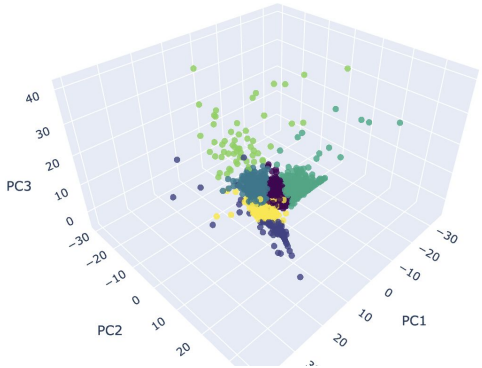
❌Mauvaise répartition
❌Silhouette score bas

Agglomerative C.

- Silhouette: 0.277
- Répartition: 252, 52, 657, 1845, 936, 1258
- Qualité: Séparation correcte

❌Meilleure séparation possible





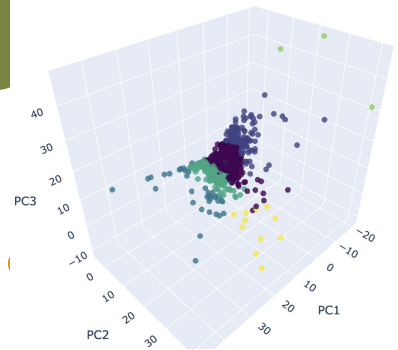
KMeans

- Silhouette: 0.338
- Répartition: 1329, 115, 2189, 513, 58, 796
- Qualité: Très bonne séparation

✅Répartition correcte
✅Meilleur silhouette score

Méthode sélectionnée: KMeans

Usages des consommateurs



Birch

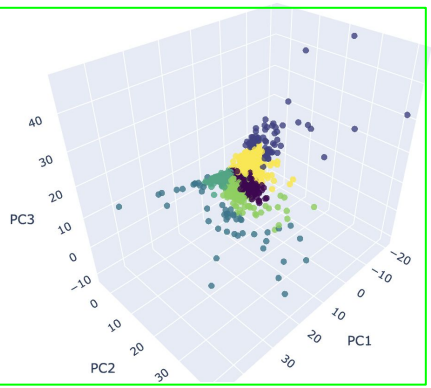
- Silhouette: 0.25
- Répartition: 3147, 123, 38, 1678, 3, 11
- Qualité: Séparation correcte

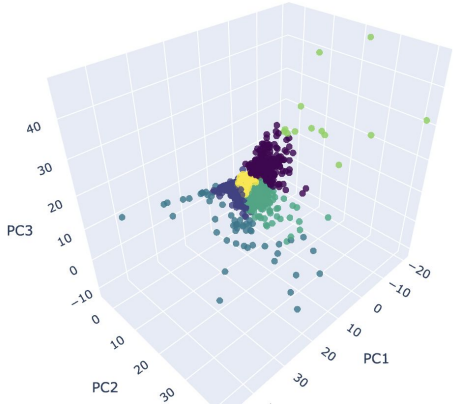
❌Très mauvaise répartition
❌Meilleure séparation possible

Agglomerative C.

- Silhouette: 0.295
- Répartition: 2095, 72, 49, 848, 1568, 368
- Qualité: Bonne séparation

✅Répartition correcte
✅Meilleur silhouette score





KMeans

- Silhouette: 0.269
- Répartition: 284, 1225, 53, 1753, 13, 1672
- Qualité: Bonne séparation

✅Répartition correcte
❌Silhouette score

Méthode sélectionnée: Agglomerative Clustering

02

Personas

Personas pour chaque cluster

Attitudes: cluster 1 - Le jardinier indépendant

- S'adonne à des activités de jardinage, mais n'entretient pas de balcon ou de terrasse. Cela indique une préférence pour le jardinage traditionnel dans un espace spacieux.
- Aime avoir un grand jardin, ce qui témoigne d'un lien fort avec les activités de plein air et peut-être d'une préférence pour la vie en milieu rural ou en banlieue spacieuse.
- Engagement modéré dans des activités liées à leur espace de vie, ce qui pourrait indiquer un équilibre entre les intérêts intérieurs et extérieurs.

Attitudes: cluster 2 – Les retraités paisibles

- Entretenir un jardin sans s'occuper des balcons ou des terrasses, en suggérant de se concentrer sur les activités de jardinage au niveau du sol, ce qui convient à un espace plus petit et plus facile à gérer.
- Préfère un jardin compact, ce qui indique un désir d'espace extérieur plus facile à entretenir, adapté aux personnes à la retraite qui recherchent des passe-temps moins exigeants sur le plan physique.
- Montre un niveau d'activité faible à modéré lié à son espace de vie, ce qui peut refléter un style de vie détendu avec un engagement sélectif dans des activités de plein air.

Attitudes: cluster 3 – La famille active

- Participe activement au jardinage, montrant une préférence pour le jardinage traditionnel, probablement en tant qu'activité familiale.
- Préfère un jardin de taille moyenne, qui offre suffisamment d'espace pour les activités familiales à l'extérieur sans être trop grand à entretenir.
- Niveau élevé d'engagement dans des activités liées à l'espace de vie, indiquant un mode de vie actif qui inclut éventuellement le jardinage et les jeux avec les enfants à l'extérieur.

Attitudes: cluster 4 – Les jeunes retraités

- Jardinage et entretien d'une terrasse, ce qui suggère un intérêt général pour les espaces de vie extérieurs et la possibilité d'apprécier diverses formes d'entretien des plantes ou de relaxation en plein air.
- Le jardin de taille moyenne offre un équilibre entre l'espace pour le jardinage et la facilité d'entretien, ce qui convient à ceux qui profitent de leurs années de retraite.
- Des niveaux d'activité très élevés liés à leur espace de vie, reflétant un investissement important de temps et d'énergie dans les espaces extérieurs, ce qui peut inclure un jardinage intensif ou l'organisation de réunions de famille.

Attitudes: cluster 5 - Le jeune, adepte d'activités en plein air

- Participe à des activités de jardinage, sans s'intéresser à l'entretien des balcons ou des terrasses, ce qui suggère une préférence pour le jardinage traditionnel, peut-être comme forme de loisir ou d'intérêt pour l'environnement.
- Occupe un jardin de taille moyenne, idéal pour quelqu'un qui débute dans la vie et qui apprécie l'espace extérieur pour la détente ou les activités sociales.
- Engagement modéré dans leur espace de vie, indiquant une approche équilibrée des activités intérieures et extérieures, avec un penchant possible pour les réunions sociales ou les passe-temps personnels dans le jardin.

Attitudes: cluster 6 – La famille minimaliste

- Montre de l'intérêt pour le jardinage mais n'entretient pas de balcon ou de terrasse, ce qui indique une préférence pour un jardinage simple, au niveau du sol, adapté à une vie de famille occupée.
- L'engagement dans les activités liées à l'espace de vie est très faible à faible, ce qui suggère soit un mode de vie occupé avec peu de temps pour les activités extérieures, soit une préférence pour les activités à l'intérieur.

Usages: cluster 1 - Les banlieusards

- S'occupe de l'entretien du jardin, ce qui indique un intérêt pour les activités de plein air et peut-être le jardinage comme passe-temps.

- Il est probable qu'ils apprécient un espace extérieur bien entretenu pour la détente et les loisirs.

Usages: cluster 2 - Les jeunes familles

- Avec un mélange de jardins et de familles plus nombreuses, l'accent est mis sur les activités extérieures qui permettent les réunions de famille, les aires de jeux pour les enfants, ou peut-être le jardinage en tant qu'activité familiale

- La priorité est donnée à l'espace extérieur pour son utilité et comme moyen de resserrer les liens familiaux.

Usages: cluster 3 – Les solitaires urbains

- Les balcons étant leur principal espace extérieur, leurs activités peuvent se concentrer sur le jardinage, les plantes décoratives ou la création d'une petite retraite en plein air.

- Ils apprécient probablement l'espace extérieur comme une oasis personnelle ou pour recevoir un petit nombre d'invités.

Usages: cluster 4 - Les jeunes indépendants

- Malgré leur jeune âge, les jeunes s'intéressent à l'entretien d'un jardin, éventuellement pour des réunions sociales ou des loisirs personnels.

- Les activités de plein air pourraient inclure un mélange d'événements sociaux et d'intérêt personnel pour le jardinage ou les projets de bricolage.

Usages: cluster 5 - Les couples retraités

- Intérêt pour l'entretien d'un jardin, suggestions d'activités autour du jardinage ou pour profiter d'un espace extérieur calme et bien entretenu.

- L'espace extérieur peut être apprécié pour sa tranquillité, sa beauté et en tant que hobby.

Usages: cluster 6 - Les familles recomposées

- Lorsque les ménages sont plus nombreux, l'accent peut être mis sur l'entretien d'un jardin pour les activités familiales, y compris le jardinage, les repas en plein air et les espaces de jeu pour les enfants.

- Le jardin sert probablement à de multiples usages, depuis les loisirs et la détente jusqu'aux rassemblements familiaux et aux divertissements.

03

Golden questions

Réaffectation en fonction du nombre de questions posées

Méthode de réaffectation par golden questions

L'objectif de cette partie est de réaffecter un consommateur dans le bon cluster en lui posant le moins de questions possible. C'est une tâche de classification, nous allons donc utiliser l'une des méthodes state of the art en machine learning: XGBoost.

Voici les étapes à réaliser:

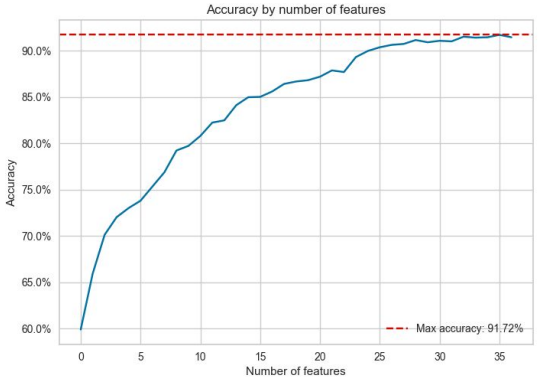
1. Entraîner un modèle XGBoost sur nos données et les cluster trouvés
2. Trouver les features (réponses aux questions) les plus importantes pour le modèle
3. Pour *i* de 1 au nombre de questions:
    a. Ré-entraîner un modèle XGBoost sur les *i* features les plus importantes
    b. Enregistrer la précision du modèle
4. Afficher la courbe des précisions en fonction du nombre de question

Ces étapes sont réalisées pour les attitudes et les usages. Les résultats sont présents sur la slide suivante

Golden questions: résultats

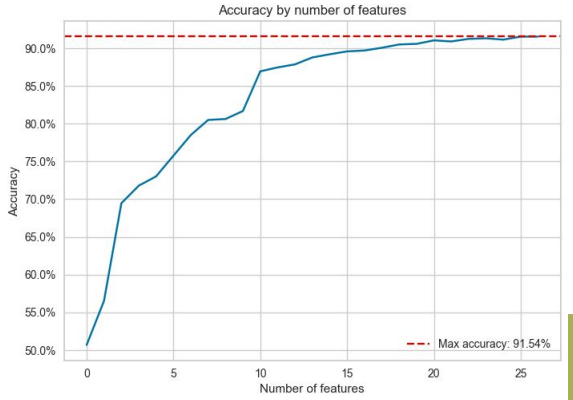
Attitudes

- Précision: ~91%
- 3 features les plus importantes:
  1. *A11_10_slice*: Un moyen de réaliser jusqu'au bout quelque chose de vos propres mains
  2. *A10_4_slice*: Les espaces extérieurs permettent d'accéder à davantage d'activités de loisirs
  3. *A11_5_slice*: Un moyen de se ressourcer, de refaire le plein d'énergie
- Résulats: À partir de 10 questions, on est capable de réaffecter 80% des consommateurs



Usages

- Précision: ~92%
- 3 features les plus importantes:
  1. *A12*: Possession d'un balcon
  2. *B4*: Fréquence des récoltes du jardin
  3. *C1_2_slice*: Fréquence consultation sites dédiés au jardinage Des sites Internet des enseignes de jardinerie
- Résulats: À partir de 10 questions, on est capable de réaffecter 87% des consommateurs



04

Variables illustratives

Réaffectation en inversant usage et attitude entre questions et clustering

Méthode de réaffectation avec inversion des questions

Cette dernière partie à pour objectif d'essayer de réaffecter les personnes dans les clusters *attitude* en utilisant les questions *usage*, et vice versa. On ajoutera également les données socio-démographiques non-utilisées jusque là

Voici les étapes à réaliser:

1. Concaténer les données attitude/usage et les données socio-démographique
2. Faire une PCA à 3 composantes sur ces nouvelles données
3. Retrouver les labels de cluster après l'ajout des données socio:
    a. Pour *attitude*, on utilise KMeans
    b. Pour *usage*, on utilise Agglomerative Clustering
4. Entraîner un classificateur (XGBoost) sur:
    a. Les données attitude + socio avec les labels *usage*
    b. Les données usage + socio avec les labels *attitude*

Ces étapes sont réalisées pour les attitudes et les usages. Les résultats sont présents sur la slide suivante

Inversion et socio: résultats

Attitudes + socio avec labels usages

- Précision: 51.3%
- Interprétation: Ceci n'est pas trop mauvais compte tenu du fait que nous avons 6 clusters donc 6 classes. Cependant, les résultats ne sont pas assez forts pour être utilisés dans un contexte professionnel. L'une des possibles explications pour ce résultat est qu'il est difficile de prédire le groupe d'attitudes d'un consommateur en fonction de ses usages. En effet, les usages sont des actions concrètes et mesurables, alors que les attitudes sont des états d'esprit et des opinions qui sont plus difficiles à mesurer.
- Matrice de confusion:

Usages + socio avec labels attitudes

- Précision: 86%
- Interprétation: Ceci est un résultat très satisfaisant et montre que les attitudes des consommateurs sont un bon indicateur des usages de leur jardin. Ce résultat est cohérent avec la partie précédente, où nous avons remarqué qu'avec le même nombre de questions, les résultats de réaffectation étaient meilleurs pour les usages, car sûrement plus simple à représenter

- Matrice de confusion: