Rapport du TP - KANTAR

Anthony BERNARD

Junyi LI

Raphael MOUROT-PELADE

1/ Clustering des données fournies

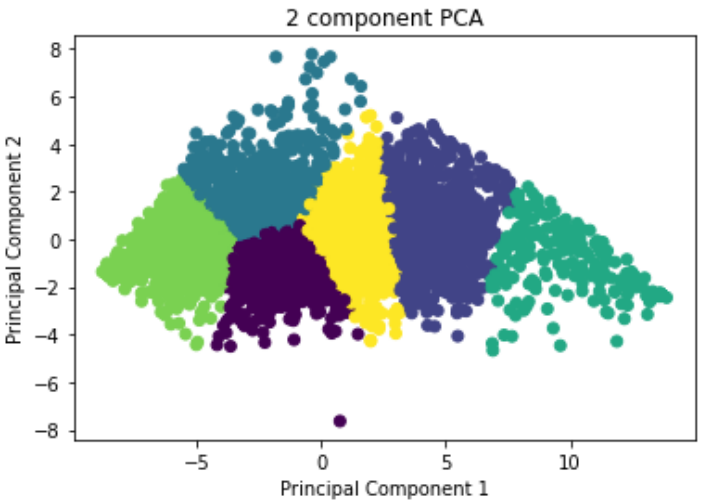
Dans cette analyse, nous avons effectué une clusterisation des 5000 individus en utilisant deux ensembles de variables. Le premier ensemble(orange) comprenait les variables A9, A10 et A11, tandis que le second ensemble(vert) comprenait un plus grand nombre de variables, allant de A11 à C1_9_slice.

Dans le cadre de notre analyse, nous avons d'abord évalué différentes méthodes de clusterisation sur l'ensemble de variables orange. Les méthodes évaluées comprenaient KMeans, la hiérarchie (représentée par un dendrogramme) et t-SNE. Nous avons choisi d'effectuer une séparation en 6 clusters car elle présentait un bon ratio variance intra-groupe/variance inter-groupe tout en ayant une diversité de profils.
Voici un résumé de nos observations pour chaque méthode :

KMeans

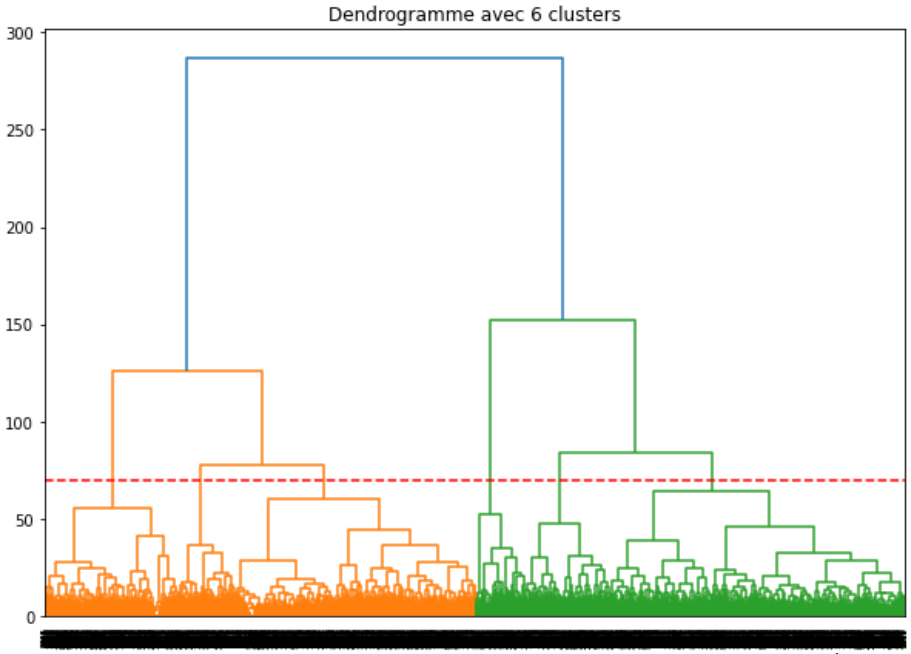
KMeans est un algorithme de clustering non supervisé qui partitionne les données en K clusters distincts, où chaque observation appartient au cluster avec la moyenne la plus proche.

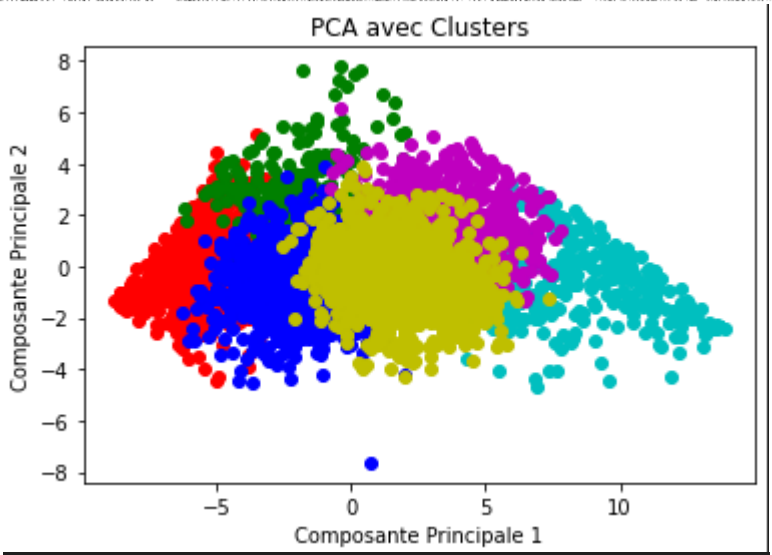
```
Variance intra-groupe pour le cluster 1: 0.6452536108409659
Variance intra-groupe pour le cluster 2: 0.8711214998263773
Variance intra-groupe pour le cluster 3: 0.9138471512729727
Variance intra-groupe pour le cluster 4: 1.038052263369652
Variance intra-groupe pour le cluster 5: 0.7117490656003871
Variance intra-groupe pour le cluster 6: 0.7448781807397619
Variance inter-groupe : 4.306536904246425
```



Dendrogramme

Le dendrogramme est une représentation visuelle d'un clustering hiérarchique. Il permet de visualiser la formation des clusters et peut aider à déterminer le nombre optimal de clusters.

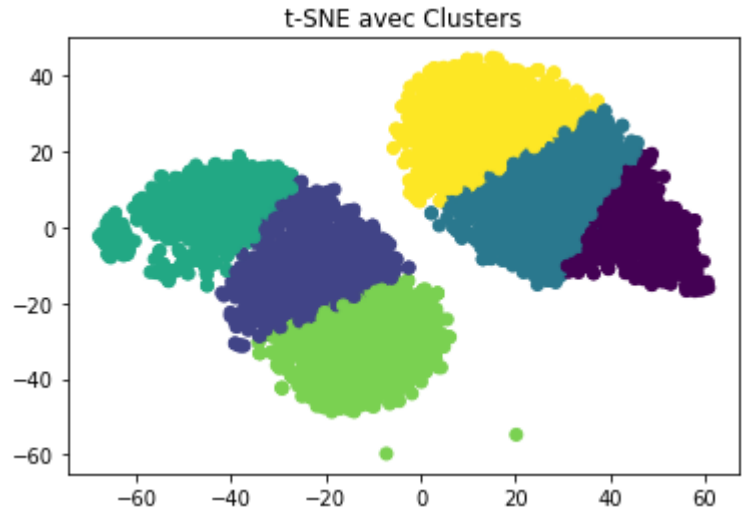





```
Variance intra-groupe pour le cluster 1: 0.860892765962681
Variance intra-groupe pour le cluster 2: 0.8845579960407182
Variance intra-groupe pour le cluster 3: 0.7253903590615347
Variance intra-groupe pour le cluster 4: 0.7777789392113447
Variance intra-groupe pour le cluster 5: 1.0174571828624852
Variance intra-groupe pour le cluster 6: 0.7019445121097743
Variance inter-groupe : 4.401876605024499
```

t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) est une technique de réduction de dimensionnalité qui est particulièrement bien adaptée pour la visualisation de jeux de données de grande dimension. Elle permet de visualiser les clusters en deux dimensions, ce qui peut aider à comprendre la structure des données.




```
Variance intra-groupe pour le cluster 1: 4.4928882367537595
Variance intra-groupe pour le cluster 2: 4.760056146693664
Variance intra-groupe pour le cluster 3: 5.345093665155894
Variance intra-groupe pour le cluster 4: 5.282649129903465
Variance intra-groupe pour le cluster 5: 4.535721825985197
Variance intra-groupe pour le cluster 6: 4.931512912238625
Variance inter-groupe : 27.034244932009997
```

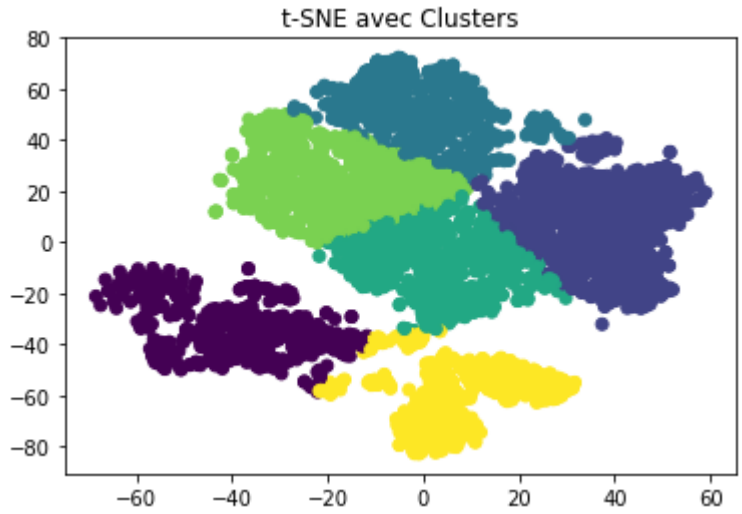
Profilage des groupes

Dans notre cas, après avoir évalué ces méthodes, nous avons décidé d'utiliser la méthode t-SNE, car elle présentait le meilleur ratio sur ses variances inter et intra groupes.
On peut alors réaliser un profilage des différents groupes en observant les réponses fournies par les différents groupes. Notre critère de sélection de réponses caractéristiques sera la majorité absolue.

Groupe 1 : Moins engagé, apprécie la nature, valorise le loisir en plein air, reconnaît l'importance du bien-être extérieur.
Groupe 2 : Apprécie le jardinage, valorise le contact avec la nature, conscient des contraintes, trouve un équilibre entre les bénéfices et les efforts.
Groupe 3 : Apprécie les avantages des espaces extérieurs, moins investi dans l'aménagement, valorise le contact avec la nature, pas particulièrement intéressé par l'apprentissage.
Groupe 4 : Très engagé, passionné par l'aménagement extérieur, valorise le contact avec la nature, conscient de l'importance de la préservation de l'environnement.
Groupe 5 : Attitude modérée, s'intéresse à l'aménagement, apprécie le jardinage, voit les espaces extérieurs comme des lieux de loisirs.
Groupe 6 : Attitude modérée, s'intéresse à l'aménagement, apprécie le jardinage, considère les espaces extérieurs comme une source de contraintes, valorise le contact avec la nature.

Ensemble vert

Nous avons ensuite adopté la même méthode pour la sélection verte, on obtient également la clusterisation la plus satisfaisante avec le t-SNE en 6 groupes.




```
Variance intra-groupe pour le cluster 1: 6.753482928794227
Variance intra-groupe pour le cluster 2: 7.446867728233212
Variance intra-groupe pour le cluster 3: 6.365454355094967
Variance intra-groupe pour le cluster 4: 5.300641472921522
Variance intra-groupe pour le cluster 5: 6.118917994678167
Variance intra-groupe pour le cluster 6: 6.421485367040931
Variance inter-groupe : 31.348335010526498
```

On peut alors dresser un profilage des groupes ainsi obtenus :

Groupe 1 : Espaces extérieurs limités, très actifs en ligne, recherchent des conseils sur le jardinage.
Groupe 2 : Grands espaces extérieurs, très actifs en ligne, recherchent des conseils sur le jardinage.
Groupe 3 : Ont un jardin, pas d'échanges d'outils ou de graines, consultent fréquemment des sites de jardinage.
Groupe 4 : Ont un jardin, prêtent rarement du gros outillage, consultent moins souvent des blogs et sites d'experts.
Groupe 5 : Ont un jardin, empruntent et prêtent rarement de l'outillage, consultent fréquemment des sites de jardinage.
Groupe 6 : Ont une terrasse, pas d'échanges d'outils ou de graines, très actifs en ligne, recherchent des conseils sur l'aménagement des espaces extérieurs.

Conclusion

En conclusion, notre analyse de clustering a permis d'identifier des groupes distincts d'individus en fonction de leurs réponses à un ensemble de variables. Ces groupes peuvent être utilisés pour mieux comprendre les tendances et les schémas dans les données.

2/ Réaffectation des individus dans les groupes avec variables actives

Dans le cadre de notre analyse, nous avons cherché à établir un algorithme permettant de réaffecter les individus dans les groupes. Pour ce faire, nous avons utilisé les variables utilisées pour construire les clusters et nous avons formé deux modèles de classification supervisée : Random Forest et XGBoost.

Dans le but de maximiser le pourcentage de bon classement tout en minimisant le nombre de variables utilisées, nous avons effectué une sélection de caractéristiques. Cette étape nous a permis d'identifier les "Golden Questions", c'est-à-dire les variables les plus informatives pour prédire le groupe d'un individu. Nous avons choisi de regarder les 10 questions les plus importantes.

L'objectif final de notre analyse était de déterminer un algorithme qui pourrait être utilisé sur une autre enquête. Ainsi, de nouveaux individus pourront répondre aux "Golden Questions" et être affectés dans les différents groupes. Les modèles de classification que nous avons formés (Random Forest et XGBoost) peuvent être utilisés à cette fin, car ils peuvent faire des prédictions sur de nouvelles données en se basant sur les réponses aux "Golden Questions".

Random Forest

Le modèle Random Forest est un algorithme d'apprentissage ensembliste qui combine les prédictions de plusieurs arbres de décision. Chaque arbre est formé sur un sous-ensemble aléatoire des données d'apprentissage, ce qui permet au modèle de capturer une variété de tendances dans les données.

XGBoost

XGBoost (eXtreme Gradient Boosting) est un autre algorithme d'apprentissage ensembliste qui utilise le boosting pour combiner les prédictions de plusieurs modèles plus simples. Contrairement à Random Forest, qui forme chaque arbre indépendamment, XGBoost forme chaque nouveau modèle pour corriger les erreurs commises par l'ensemble des modèles précédents.

Après avoir entraîné et évalué les deux modèles, nous avons comparé leurs performances pour déterminer lequel était le plus précis pour prédire le groupe d'un nouvel individu. Nous avons utilisé l'exactitude de la classification comme métrique de performance.

Pourcentage de bon classement	Sélection Orange	Sélection Verte
Random Forest	57,6%	64,2%
XGBoost	62,9%	67,5%

Conclusion

En conclusion, notre analyse a montré que XGBoost fournissait de bons résultats pour la réaffectation d'individus dans les groupes formés précédemment.