KANTAR

Alexandre Bailly | Lina Farchado | Florian Tigoulet

0.DATASET

Les variables vertes se concentrent sur les caractéristiques des espaces extérieurs, leur entretien, leur utilisation, ainsi que les comportements liés à l'échange, au prêt ou à la consultation de ressources concernant le jardinage et l'aménagement extérieur. Elles reflètent des pratiques concrètes et mesurables, comme le temps passé, la taille des espaces, ou les fréquences d'activités.

Les variables oranges, quant à elles, explorent les perceptions, attitudes et valeurs personnelles associées aux espaces extérieurs. Elles abordent l'importance accordée au jardinage, la décoration, les aspects utilitaires ou écologiques, ainsi que les émotions ou bienfaits liés à ces espaces, offrant une perspective plus subjective et affective.

1.CLUSTERING

Objectif:

L'objectif est de trouver les meilleures méthodes de clusterisation pour trouver les meilleurs résultats sur les variables vertes et oranges.

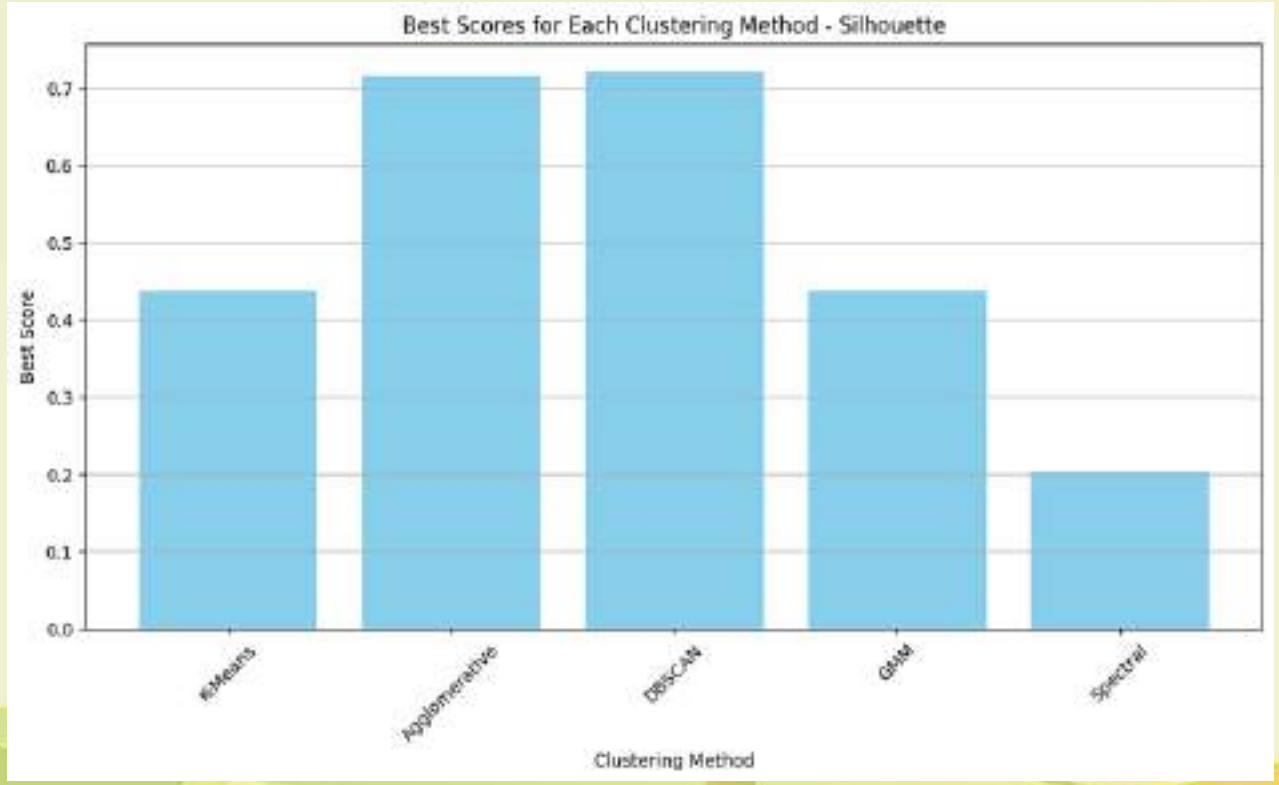
Méthodologie:

On va d'abord faire un benchmark sur 5 différents algorithmes de clusterisation :

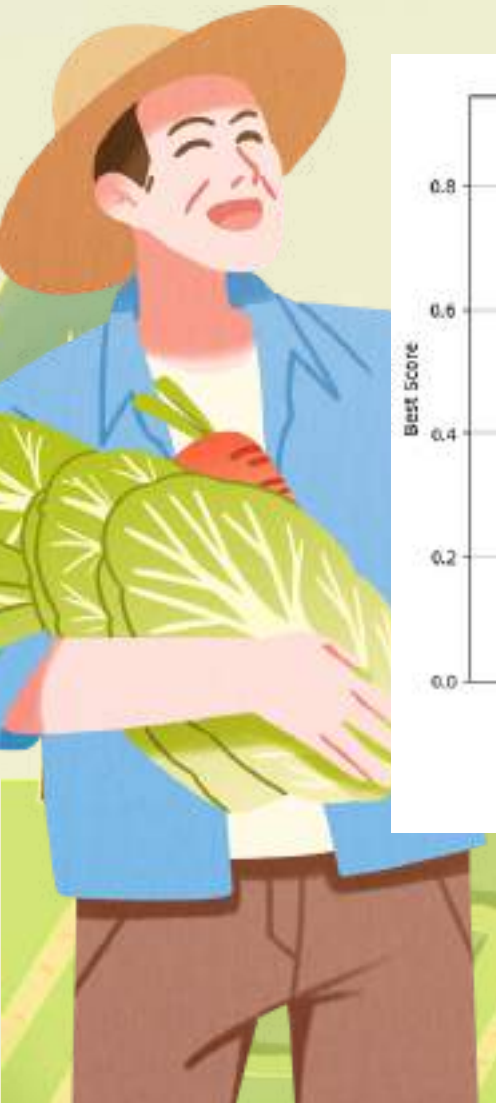
- K-Means
- AgglomerativeClustering
- DBSCAN
- Gaussian Mixture Model
- Spectral Clustering

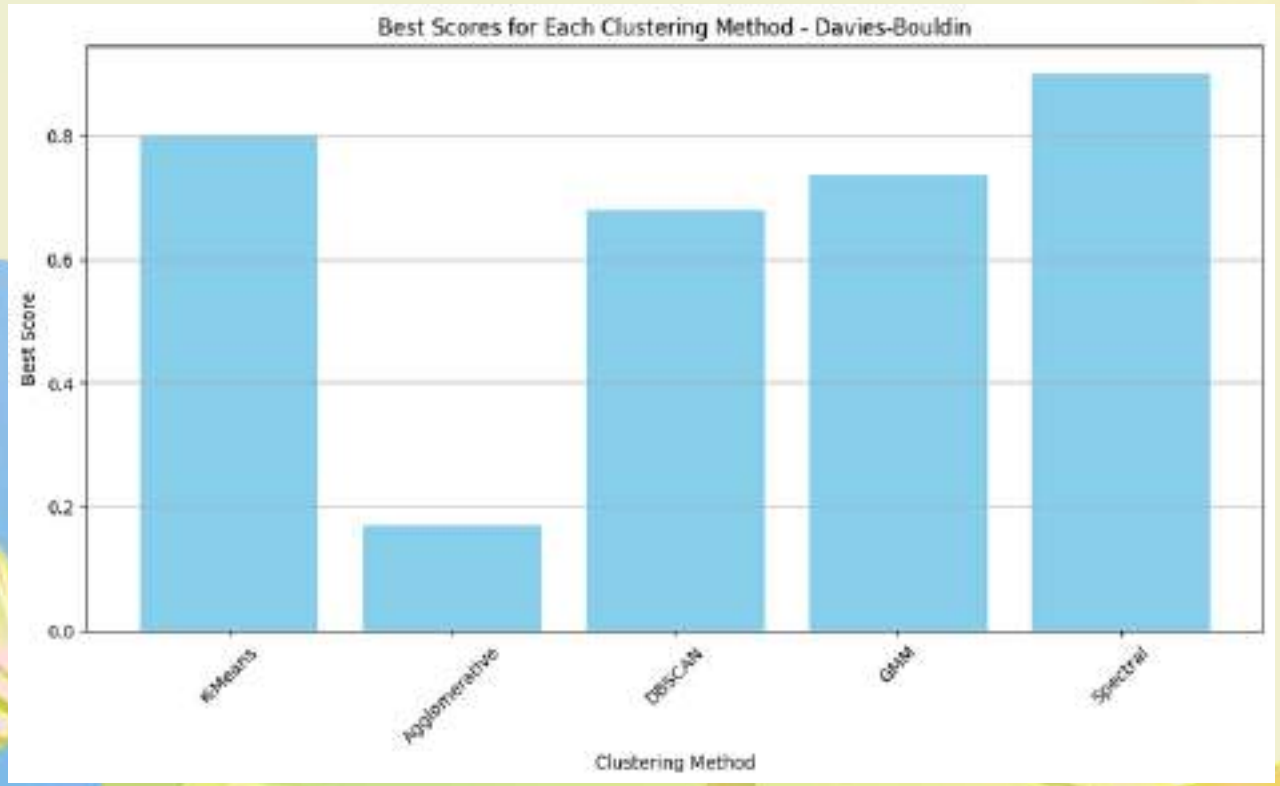
Ce benchmark a pour but de trouver les meilleurs hyperparamètres (nombre de clusters, pca components), pour trouver la meilleure méthode de clusterisation.
Ensuite, pour chaque variable, calculer les variances intra-groupes, inter-groupes et ratio.

BENCHMARK (VERT) SILHOUETTE

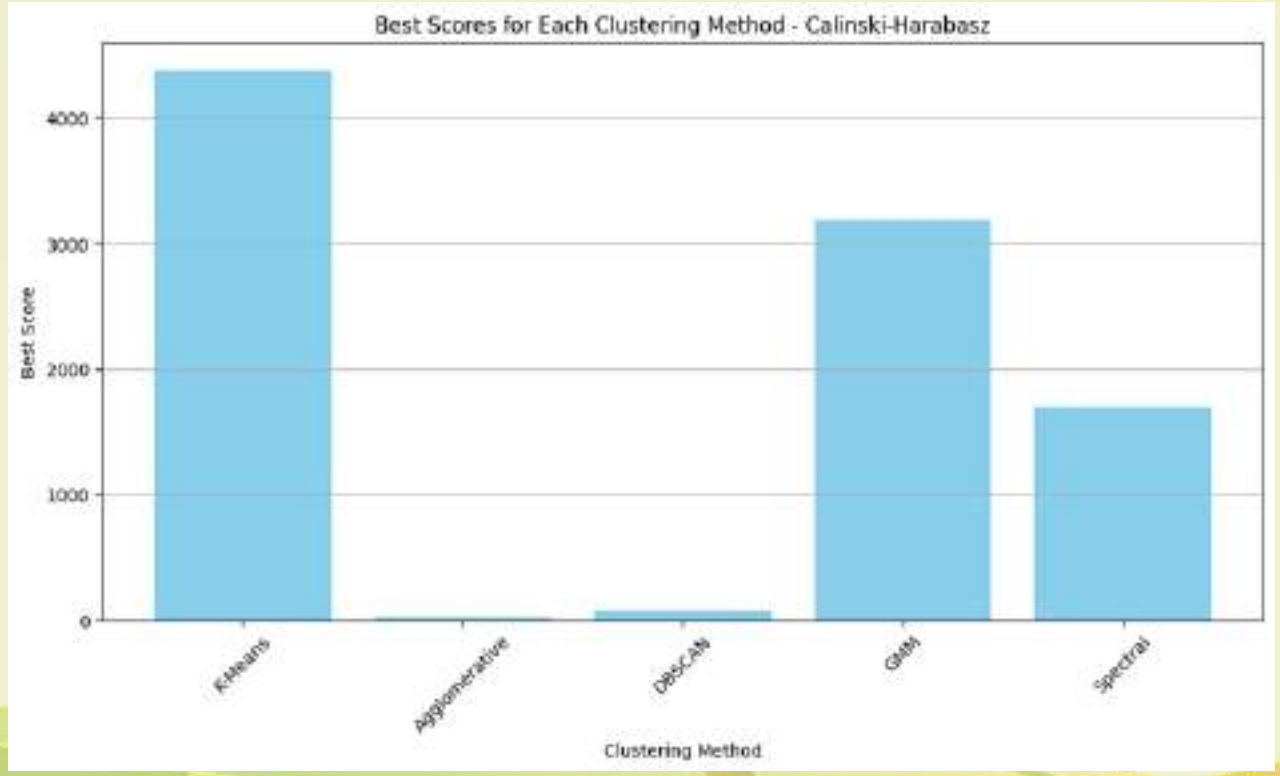


BENCHMARK (VERT) DAVIES-BOULDIN





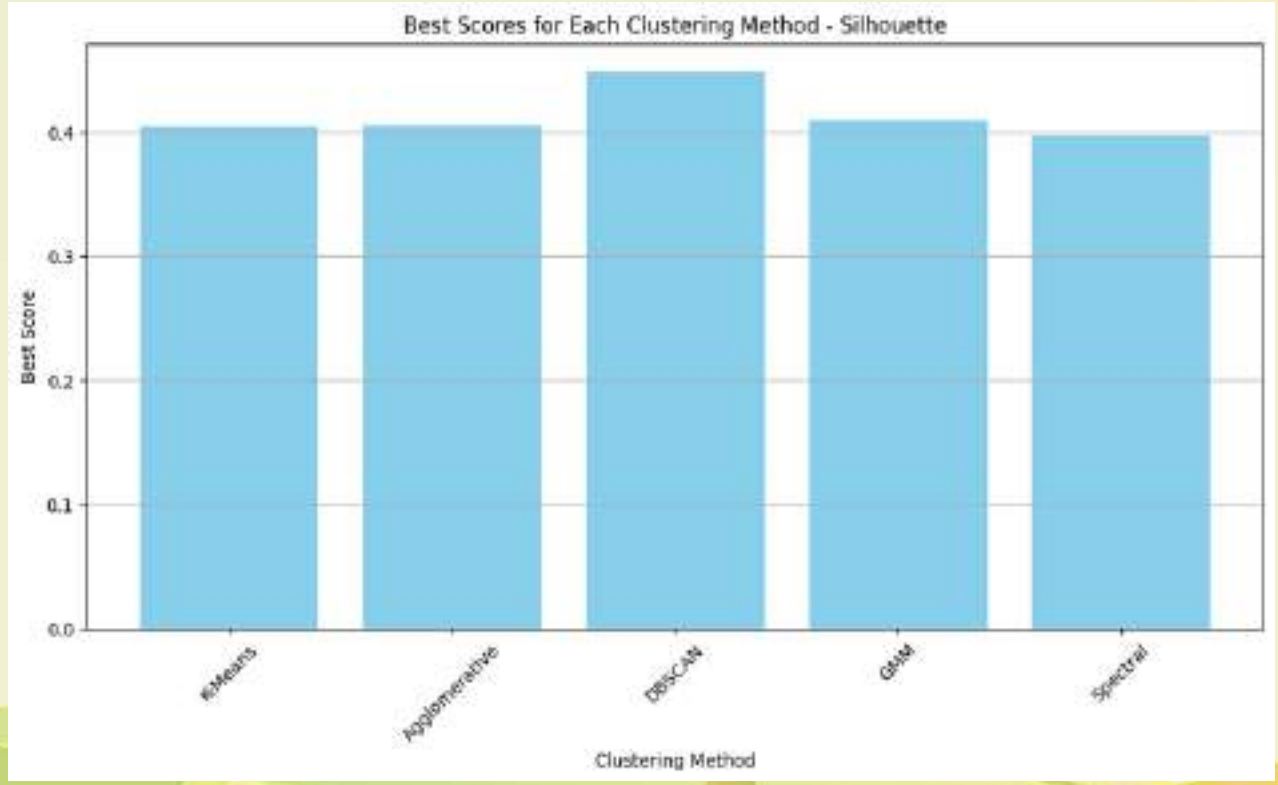
BENCHMARK (VERT) CALINSKI-HARABASZ



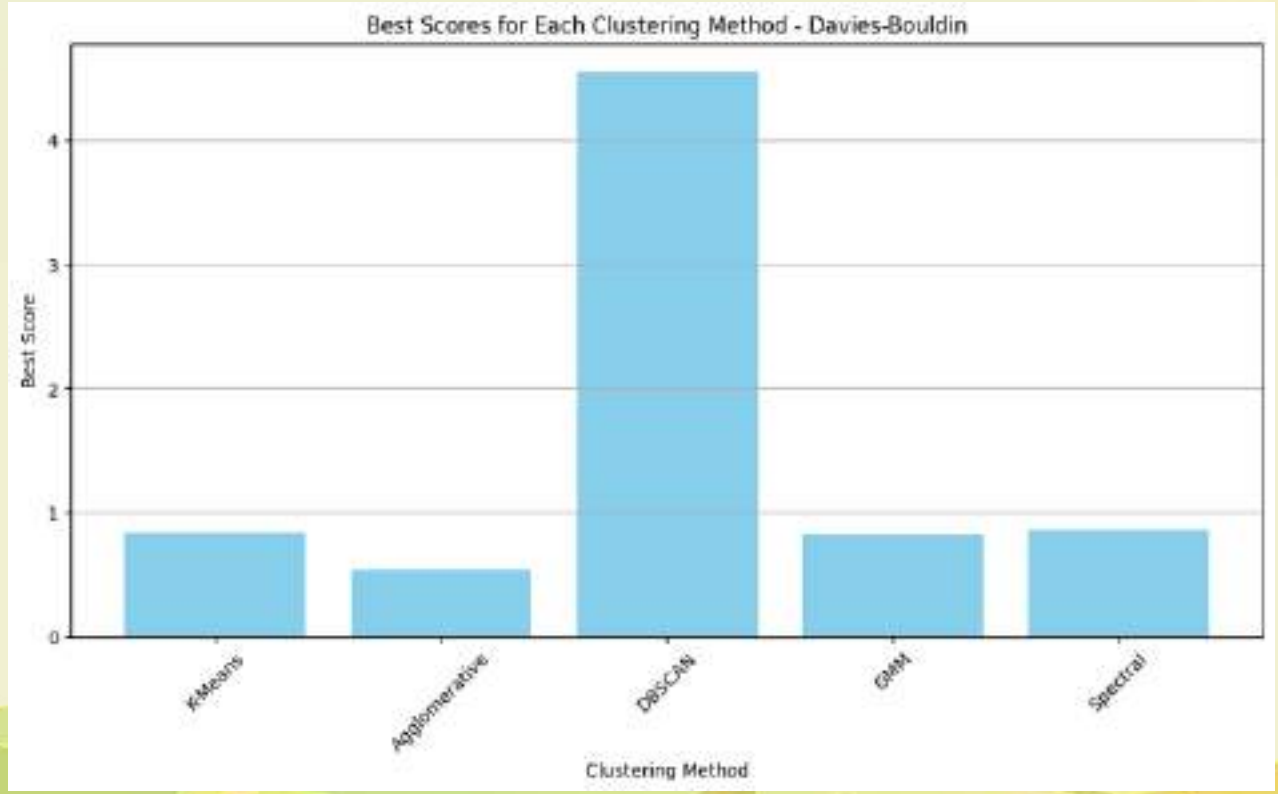
BENCHMARK (VERT)

En regardant les graphes et en analysant les résultats, on se rend compte que le K-means est celui qui a, en moyenne, les meilleurs résultats.
On peut le voir en regardant par exemple le DBSCAN, qui est très performant pour la Sihouette mais laisse à désirer pour les 2 autres metriques.

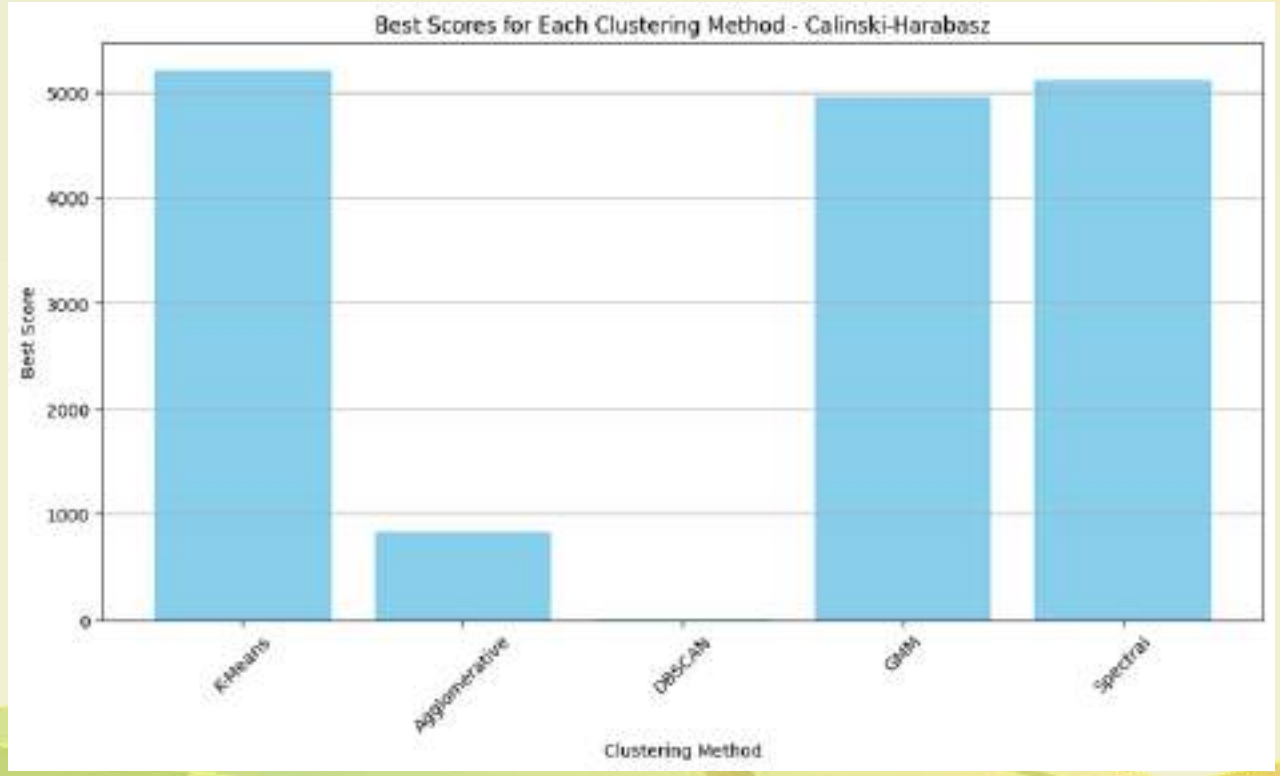
BENCHMARK (ORANGE) SILHOUETTE



BENCHMARK (ORANGE) DAVIES-BOULDIN



BENCHMARK (ORANGE) CALINSKI-HARABASZ



BENCHMARK (ORANGE)

Pour les mêmes raisons que pour le benchmark vert, ici nous avons choisi le K-Means, qui a les meilleurs scores dans chaque métrique, en moyenne.

CALCUL DES RATIOS

Vert

Variance Intra-Groupes:
- 2.126

Variance Inter-Groupes:
- 3.322

Ratio:
- 1.562

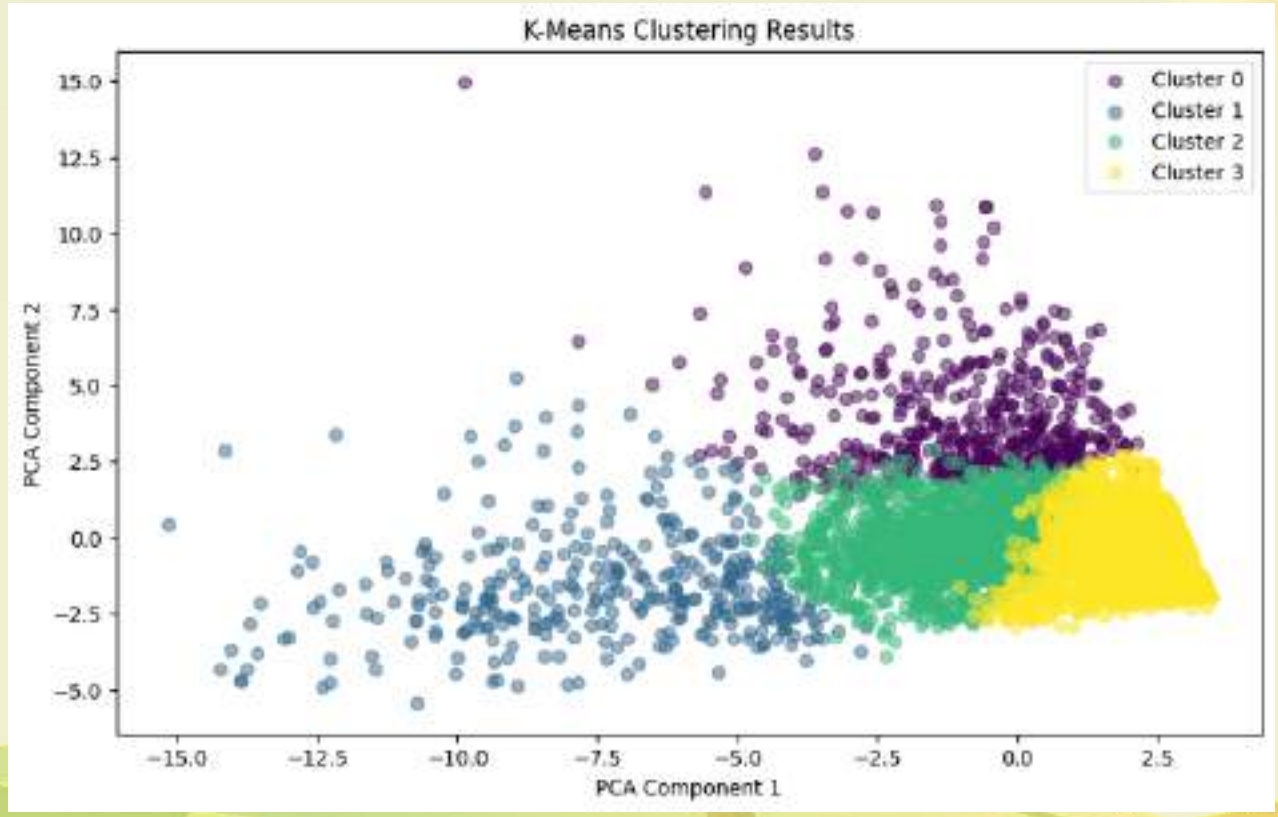
Orange

Variance Intra-Groupes:
- 2.239

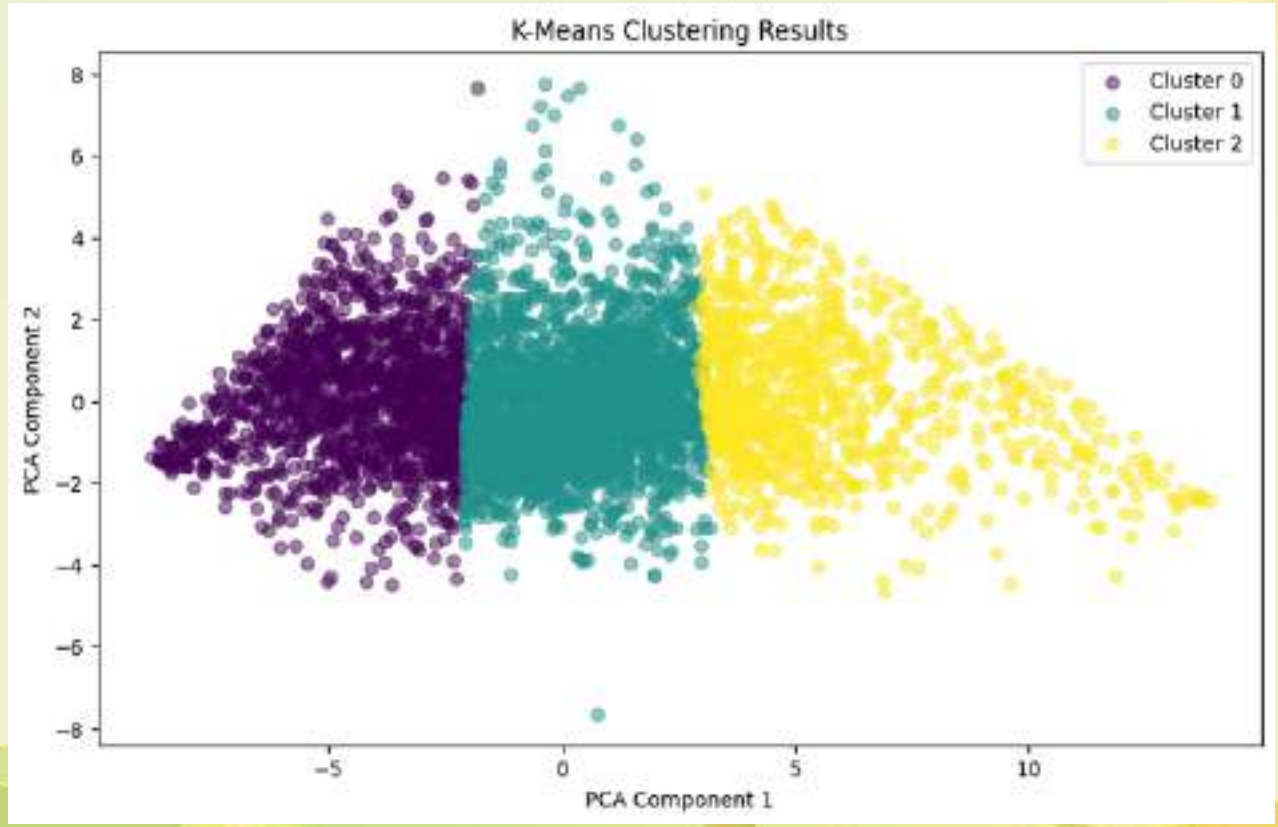
Variance Inter-Groupes:
- 3.476

Ratio:
- 1.552

VISUALIZATIONS DES CLUSTERS (VERT)



VISUALIZATIONS DES CLUSTERS (ORANGE)



EXPLICATION DES CLUSTERS (VERT)

Cluster 0

Ce cluster regroupe des individus qui :

- Consacrent un temps significatif à l'entretien de leurs espaces extérieurs, surtout au printemps et en été, mais beaucoup moins en automne et hiver.
- Ne s'intéressent pas à des contenus numériques liés au jardinage sur Instagram ou n'en consultent jamais.

EXPLICATION DES CLUSTERS (VERT)

Cluster 1

Ce cluster regroupe des individus qui :

- Passent très peu de temps à l'entretien de leurs espaces extérieurs, quelle que soit la saison.
- Consultent régulièrement des profils Instagram dédiés au jardinage.
- Lisent des blogs d'experts en aménagement quelques fois par mois pour s'inspirer ou s'informer.

EXPLICATION DES CLUSTERS (VERT)

Cluster 2

Ce cluster regroupe des individus qui :

- Passent très peu de temps à entretenir leurs espaces extérieurs, même en été.
- Ne consultent jamais de contenus numériques liés au jardinage ou à l'aménagement des espaces extérieurs (Instagram, Facebook, blogs, forums).
- Manifestent peu d'intérêt pour les échanges ou discussions autour de l'entretien des espaces extérieurs.

EXPLICATION DES CLUSTERS (VERT)

Cluster 3

Ce cluster regroupe des individus qui :

- Ne consultent jamais de contenus numériques en rapport avec le jardinage, que ce soit sur Instagram, Facebook, Pinterest, des blogs ou des forums.
- Peuvent être caractérisés par un désintérêt total pour les ressources numériques ou en ligne dédiées à l'aménagement ou à l'entretien des espaces extérieurs.

EXPLICATION DES CLUSTERS (ORANGE)

Cluster 0

Ce cluster regroupe des individus qui :

- Ne perçoivent pas les espaces extérieurs comme étant des sources de contraintes.
- Préfèrent un entretien modéré des espaces extérieurs, plutôt que des espaces totalement sauvages.
- Investissent raisonnablement dans l'aménagement de leurs espaces extérieurs.
- Considèrent les espaces extérieurs comme utilitaires et recherchent parfois des informations en ligne pour les entretenir ou les aménager.

EXPLICATION DES CLUSTERS (ORANGE)

Cluster 1

Ce cluster regroupe des individus qui :

- Investissent peu dans l'aménagement et l'entretien de leurs espaces extérieurs.
- Sont peu intéressés par les nouveautés liées à l'aménagement des espaces extérieurs ou par la recherche d'informations sur le sujet.
- Ne préfèrent pas particulièrement des espaces extérieurs sauvages ni trop entretenus.
- Ne consomment pas de contenus numériques (vidéos, tutoriels) en rapport avec le jardinage ou l'aménagement des espaces extérieurs.

EXPLICATION DES CLUSTERS (ORANGE)

Cluster 2

Ce cluster regroupe des individus qui :

- Ont un désintérêt marqué pour les espaces extérieurs, ne s'impliquant pas dans leur aménagement ou leur entretien.
- N'investissent pas financièrement dans leurs espaces extérieurs.
- Ne recherchent ni d'informations ni de tutoriels concernant le jardinage ou l'entretien des espaces extérieurs.
- Ne s'intéressent pas aux nouveautés liées aux espaces extérieurs.

2. RÉAFFECTATION AVEC VARIABLES ACTIVES

Objectif:

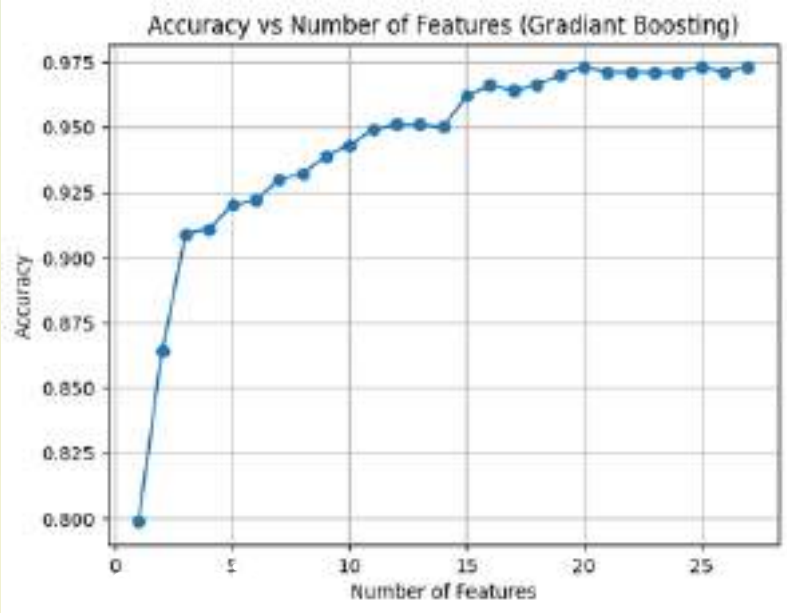
On cherche à développer un algorithme de classification permettant de réaffecter les individus dans leurs clusters respectifs à partir des variables actives out en minimisant le nombre de questions nécessaires.

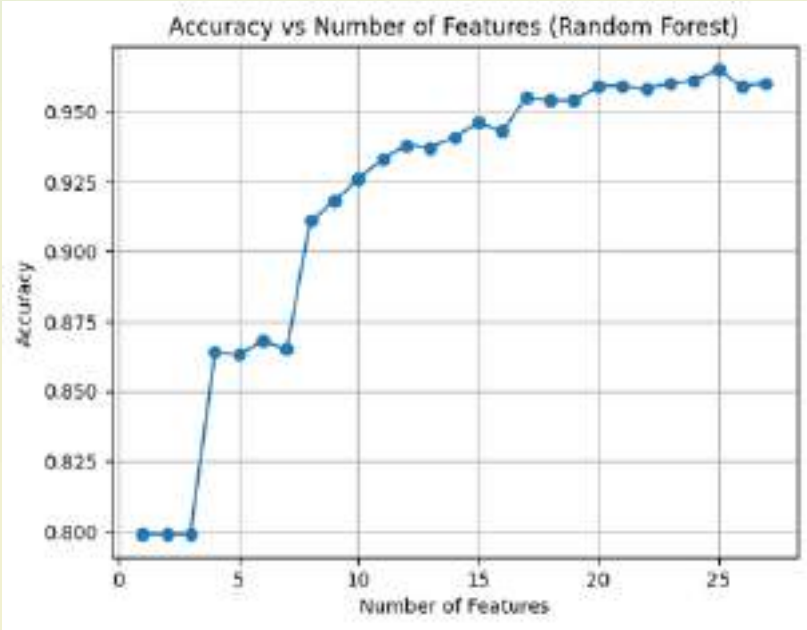
Méthodologie:

On a donc utilisé deux algorithmes principaux, Random Forest et Gradient Boosting, pour évaluer leurs performances sur deux types de variables.
Pour chaque type de variable (vert et orange), nous avons :

1. Entraîné les deux models.
2. Identifié l'ordre d'importance des questions grâce à l'analyse des importances des variables fournies par les modèles.
3. Itérativement réduit le nombre de variables (questions) utilisées, afin d'évaluer le meilleur compromis entre précision et simplicité pour chacun des models (On retire les questions dans l'ordre des moins importantes au plus importantes).
4. Trouver le meilleur model et la meilleur minimisation des questions.

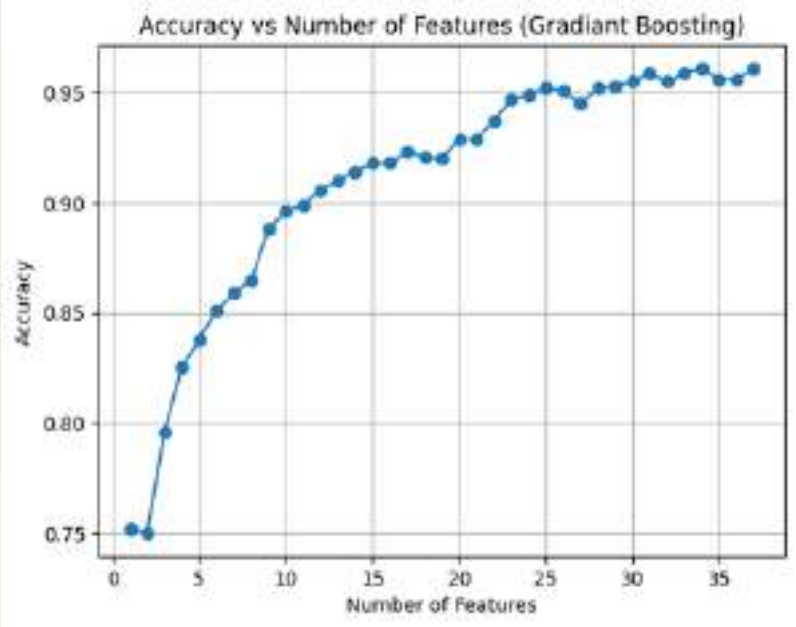
RÉSULTAT (VERT)

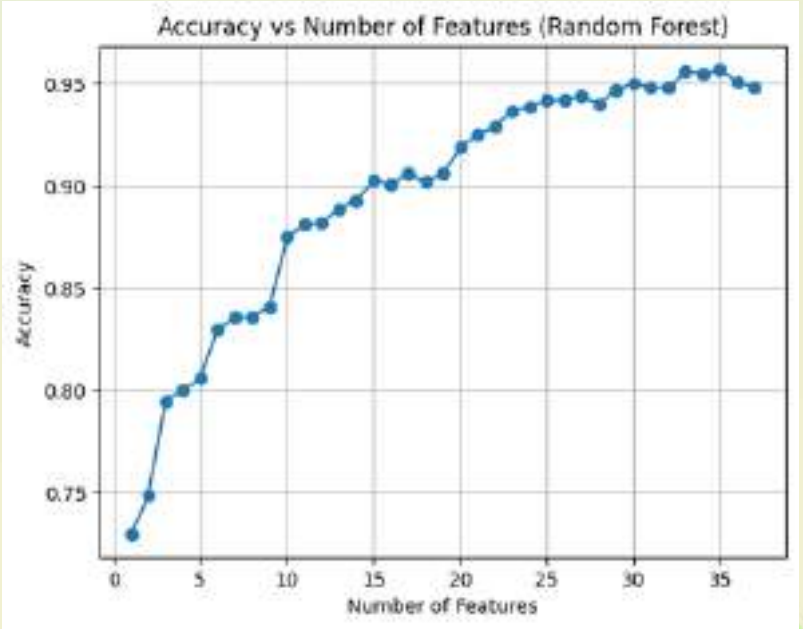




Gradient Boosting est meilleur que Random Forest pour cette tâche, offrant de meilleurs résultats. Par conséquent, le choix optimal est Gradient Boosting.
L'utilisation de seulement 15 variables permet d'atteindre une précision proche de 96.5 %, trouvant ainsi le meilleur équilibre entre performance et complexité. Au-delà de ce nombre, les améliorations de précision deviennent négligeables et le modèle commence à ralentir, ce qui fait de cette configuration le choix optimal pour maximiser les performances de classification tout en minimisant le nombre de variables.

RÉSULTAT (ORANGE)





Tout comme pour les variables verte, Gradient Boosting est meilleur que Random Forest. En utilisant 23 variables, nous atteignons une précision proche de 95 %, ce qui constitue le meilleur équilibre entre performance et le nombre de variables.

CONCLUSION

Nous avons identifié des modèles précis pour réaffecter les individus dans leurs clusters respectifs à partir des golden questions, tout en minimisant le nombre de variables nécessaires.

- Variables vertes :
  - Gradient Boosting est recommandé
  - Avec 15 variables, nous atteignons une précision proche de 96.5 %
- Variables orange :
  - Gradient Boosting est recommandé
  - Avec 23 variables, la précision atteint près de 95 %

Ces résultats démontrent que les algorithmes développés sont adaptés et performants pour des enquêtes futures, offrant une réaffectation précise des individus tout en réduisant au maximum le nombre de questions nécessaires.

3. RÉAFFECTATION AVEC VARIABLES ILLUSTRATIVES

Objectif:

L'objectif est de déterminer des algorithmes robustes et généralisables qui pourront être utilisés sur un fichier annexe ou un fichier client, où seules les variables sélectionnées sont disponibles. Ces algorithmes doivent permettre de réaffecter les individus dans leurs groupes respectifs de manière précise et efficace.

Méthodologie:

Deux algorithmes principaux (Random Forest et Gradient Boosting), ont sélectionnés pour leur capacité à gérer des données complexes et fournir des performances élevées.
On les a testés sur deux types de variables illustratives :

- Pour la segmentation basée sur les variables Vertes, nous avons utilisé les variables Oranges et une liste spécifique de variables donné dans l'énoncé.
- Pour la segmentation basée sur les variables Oranges, nous avons utilisé les variables Vertes et cette même liste spécifique de variables.

Lorsque leurs performances étaient trop faibles, nous avons tenté de les améliorer en réalisant une optimisation des hyperparamètres grâce à une recherche en grille (GridSearchCV).

RÉSULTAT (VERT) :
AFFECTATION AVEC LES VARIABLES EN ORANGE

Classification Report (Green Variables) with Random Forest:

	precision	recall	f1-score	support
0	0.38	0.07	0.11	183
1	0.56	0.28	0.37	256
2	0.48	0.40	0.44	100
3	0.60	0.85	0.71	541
accuracy			0.58	1000
macro avg	0.49	0.40	0.41	1000
weighted avg	0.55	0.58	0.53	1000

Classification Report (Specific Variables with Gradient Boosting):

	precision	recall	f1-score	support
0	0.44	0.18	0.26	103
1	0.60	0.30	0.40	256
2	0.47	0.40	0.43	100
3	0.61	0.84	0.70	541
accuracy			0.59	1000
macro avg	0.53	0.43	0.45	1000
weighted avg	0.58	0.59	0.55	1000

Les deux modèles ont obtenu des résultats moyens et similaires. Une optimisation des hyperparamètres a ensuite été effectuée, mais celle-ci n'a pas permis d'améliorer les performances.

RÉSULTAT (VERT) :

AFFECTATION AVEC LES VARIABLES SPÉCIFIQUE

Classification Report (Specific Variables) with Random Forest:

	precision	recall	f1-score	support
0	0.22	0.34	0.27	103
1	0.64	0.67	0.65	256
2	0.18	0.10	0.13	100
3	0.67	0.64	0.66	541
accuracy			0.56	1000
macro avg	0.43	0.44	0.43	1000
weighted avg	0.57	0.56	0.56	1000

Classification Report (Specific Variables) with Gradient Boosting:

	precision	recall	f1-score	support
0	0.14	0.01	0.02	103
1	0.74	0.69	0.71	256
2	0.10	0.01	0.02	100
3	0.67	0.92	0.77	541
accuracy			0.68	1000
macro avg	0.41	0.41	0.38	1000
weighted avg	0.58	0.68	0.61	1000

optimisation des hyperparamètres

optimisation des hyperparamètres

accuracy			0.68	1000
macro avg	0.35	0.41	0.37	1000
weighted avg	0.55	0.68	0.60	1000

accuracy			0.69	1000
macro avg	0.35	0.41	0.37	1000
weighted avg	0.55	0.69	0.61	1000

Les deux modèles ont obtenu des résultats moyens. Gradient Boosting est bien mieux que Random Forest. Une optimisation des hyperparamètres a ensuite été effectuée, augmentant surtout les performances de Random Forest.

RÉSULTAT (ORANGE) :
AFFECTATION AVEC LES VARIABLES EN VERT

Classification Report (Green Variables) with Random Forest:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	262
1	0.74	0.56	0.64	184
2	0.75	0.84	0.79	554
accuracy			0.75	1000
macro avg	0.75	0.70	0.72	1000
weighted avg	0.75	0.75	0.75	1000

Classification Report (Specific Variables with Gradient Boosting):

	precision	recall	f1-score	support
0	0.75	0.68	0.71	262
1	0.73	0.62	0.67	184
2	0.75	0.82	0.79	554
accuracy			0.75	1000
macro avg	0.75	0.71	0.73	1000
weighted avg	0.75	0.75	0.75	1000

Les deux modèles ont obtenu des résultats plutôt correct. On a pas effectué d'optimisation d'hyperparamètres étant déjà satisfait des résultats.

RÉSULTAT (ORANGE) :
AFFECTATION AVEC LES VARIABLES SPÉCIFIQUE

Classification Report (Specific Variables) with Random Forest:

	precision	recall	f1-score	support
0	0.32	0.32	0.32	262
1	0.21	0.20	0.20	184
2	0.55	0.55	0.55	554
accuracy			0.43	1000
macro avg	0.36	0.36	0.36	1000
weighted avg	0.43	0.43	0.43	1000

Classification Report (Specific Variables) with Gradient Boosting:

	precision	recall	f1-score	support
0	0.32	0.05	0.09	262
1	0.34	0.07	0.12	184
2	0.55	0.91	0.69	554
accuracy			0.53	1000
macro avg	0.40	0.35	0.30	1000
weighted avg	0.45	0.53	0.43	1000

optimisation des hyperparamètres

accuracy			0.75	1000
macro avg	0.75	0.70	0.72	1000
weighted avg	0.75	0.75	0.75	1000

Les deux modèles ont initialement obtenu de mauvais résultats. Une optimisation des hyperparamètres a ensuite été réalisée. Si cela n'a pas significativement amélioré les performances pour Random Forest, cela a permis à Gradient Boosting de progresser de 22 %, rendant ce modèle performant et bien adapté à la tâche !

3. CONCLUSION

Nous pouvons en conclure:

- **Segmentation Verte avec variables Oranges :**

Les deux modèles ont des performances similaires et moyennes. Nous recommandons Random Forest pour sa rapidité, la légère amélioration de Gradient Boosting ne justifiant pas son coût supplémentaire.
Accuracy : 0.58

- **Segmentation Verte avec variables spécifiques :**

Gradient Boosting a surpassé Random Forest avec des résultats moyens jusqu'à l'optimisation d'hyperparamètre. Après celui-ci, leurs performances sont similaires. Nous recommandons donc encore Random Forest pour la même raison qu'avant.
Accuracy : 0.68

- **Segmentation Orange avec variables Vertes :**

Les performances étant correctes et similaires, nous recommandons encore Random Forest.
Accuracy : 0.75

- **Segmentation Orange avec variables spécifiques :**

Gradient Boosting, après optimisation, a progressé de 22 %, rendant ce modèle performant. Il est recommandé pour ce scénario.
Accuracy : 0.75
En résumé, Random Forest est privilégié pour sa rapidité dans les cas similaires, et Gradient Boosting est recommandé lorsqu'il offre des gains significatifs.

CODEBASE

Vous pouvez trouver notre travail sur :
https://github.com/Misklean/kantar_extraction