Rapport de TP EPITA - KANTAR

Votre Nom
votre.email@epita.fr

24 Janvier 2025

Table des matières

1 Introduction **2**
1.1 Objectif du TP 2
1.2 Approche méthodologique 2

2 Prétraitement des données **2**
2.1 Description des données 2
2.2 Nettoyage et transformation 2

3 Clustering **2**
3.1 Mises en œuvre des clusters ORANGE et VERT 2
3.1.1 Clustering sur les variables ORANGE (A9, A10, A11) 3
3.1.2 Clustering sur les variables VERT (A11, A12, A13, A14, ...) 3
3.2 Méthodes utilisées 3
3.3 Comparaison des performances 3
3.4 Visualisation des clusters 4

4 Réaffectation des individus **4**
4.1 Utilisation des variables actives 4
4.1.1 orange 4
4.1.2 vert 5
4.2 Utilisation des variables illustratives 6
4.3 Matrices de confusion 6

5 Analyse des résultats **7**

6 Conclusion **8**

1 Introduction

1.1 Objectif du TP

L'objectif de ce travail est de réaliser une analyse de segmentation des individus à partir des données fournies par KANTAR. Nous devons identifier des groupes homogènes d'individus en utilisant plusieurs méthodes de *clustering* et réaffecter de nouveaux individus en minimisant le nombre de variables utilisées.

1.2 Approche méthodologique

L'approche suivie comprend les étapes suivantes :

— Le prétraitement des données
— La réduction de dimension (**Analyse en Composantes Principales**, ACP)
— Le *clustering* avec différentes méthodes
— La réaffectation des individus avec des modèles supervisés
— L'évaluation des performances

---

2 Prétraitement des données

2.1 Description des données

Les fichiers fournis contiennent des données socio-démographiques et comportementales des individus. Deux ensembles de variables ont été identifiés :

— **Variables ORANGE** : A9, A10, A11
— **Variables VERT** : A11, A12, A13, A14, etc.

2.2 Nettoyage et transformation

Les étapes suivantes ont été réalisées :

1. **Suppression des valeurs manquantes** : Les enregistrements contenant des valeurs manquantes ont été éliminés pour garantir la qualité des analyses ultérieures.
2. **Normalisation des variables** : Les données ont été normalisées pour assurer que chaque variable contribue de manière équitable aux analyses de *clustering*.
3. **Réduction de dimension** : Une Analyse en Composantes Principales (ACP) a été appliquée afin de réduire la dimensionnalité des données tout en préservant la majorité de l'information.

---

3 Clustering

3.1 Mises en œuvre des clusters ORANGE et VERT

Nous avons réalisé deux clusterisations distinctes en fonction des variables définies :

3.1.1 Clustering sur les variables ORANGE (A9, A10, A11)

— **Objectif** : Identifier des segments d'individus selon un ensemble réduit de variables comportementales.

— **Techniques utilisées** : BIRCH, Agglomératif, K-Means

— **Comparaison des résultats** : Les résultats obtenus avec chaque méthode ont été comparés pour déterminer la méthode la plus appropriée.

3.1.2 Clustering sur les variables VERT (A11, A12, A13, A14, ...)

— **Objectif** : Créer des groupes plus affinés en exploitant un ensemble de variables plus large.

— **Techniques utilisées** : BIRCH, Agglomératif, K-Means

— **Analyse approfondie** : Les répartitions des clusters ont été analysées en détail et les résultats comparés entre les différentes méthodes.

3.2 Méthodes utilisées

Les trois méthodes de *clustering* suivantes ont été appliquées :

1. **BIRCH** (Balanced Iterative Reducing and Clustering using Hierarchies) : Une méthode hiérarchique adaptée aux grands ensembles de données.
2. **Agglomerative Clustering** : Un *clustering* hiérarchique agglomératif basé sur la proximité des points.
3. **K-Means** : Une méthode de partitionnement basée sur les distances, visant à minimiser la variance intra-cluster.

3.3 Comparaison des performances

Les performances des différentes méthodes de *clustering* ont été évaluées à l'aide du score de silhouette, qui mesure la qualité de la séparation des clusters.

Méthode	Score de silhouette (ORANGE)	Score de silhouette (VERT)
BIRCH	0.331	0.370
Agglomératif	0.320	0.357
K-Means	**0.384**	**0.406**

TABLE 1 – Comparaison des scores de silhouette des différentes méthodes de *clustering*

Conclusion : La méthode K-Means a été retenue en raison de son meilleur score de silhouette, indiquant une meilleure qualité de séparation des clusters.

3.4 Visualisation des clusters

Les clusters obtenus ont été visualisés en 3D via l'ACP, permettant une interprétation visuelle de la segmentation.

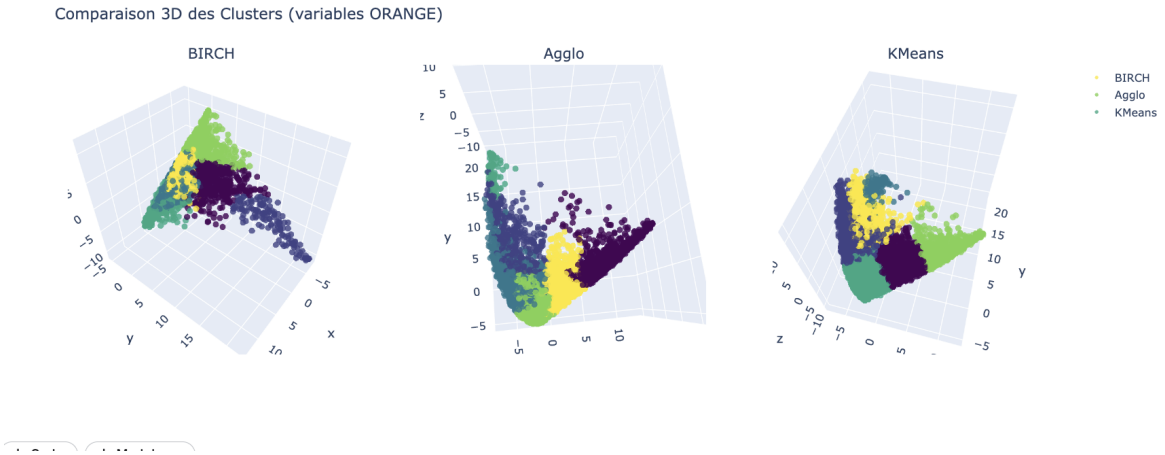


FIGURE 1 – Visualisation des clusters ORANGE

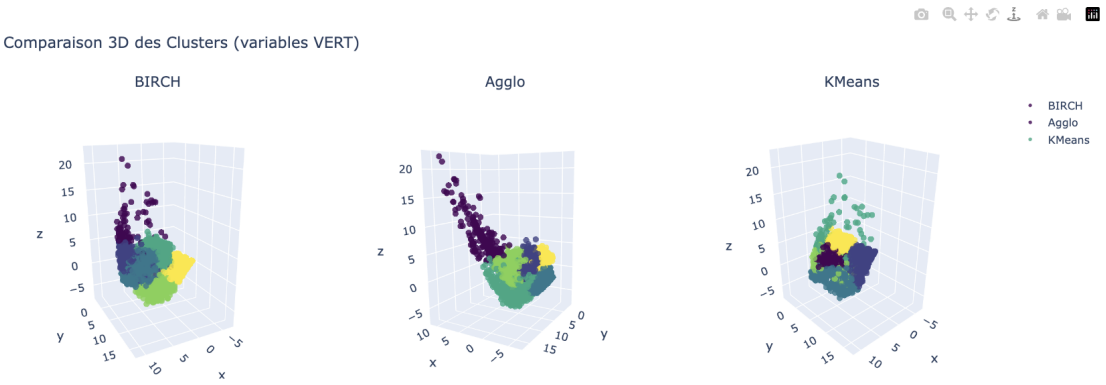


FIGURE 2 – Visualisation des clusters VERT

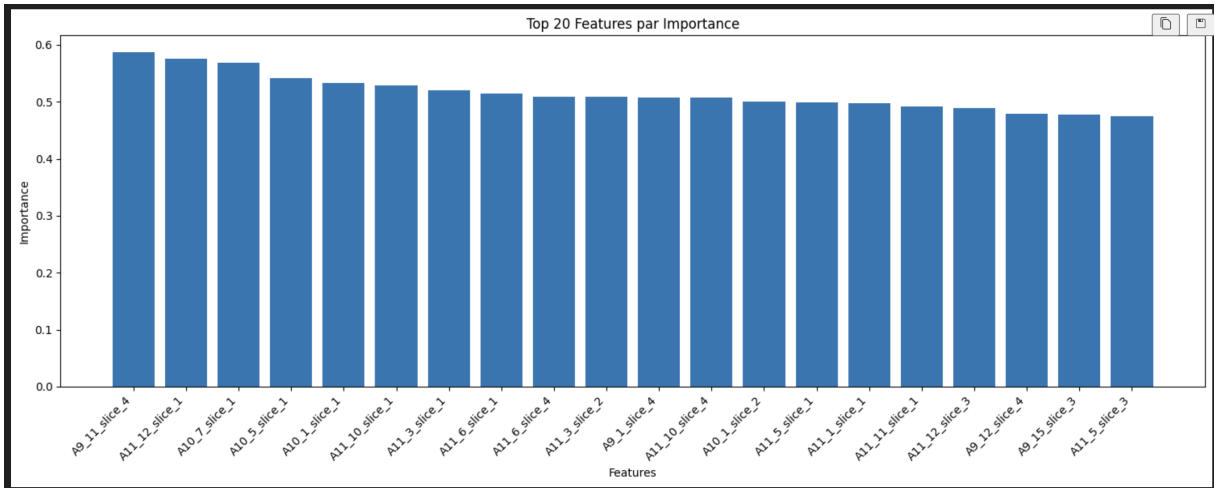
---

4 Réaffectation des individus

4.1 Utilisation des variables actives

4.1.1 orange

Une régression logistique a été utilisée pour la réaffectation des individus dans les clusters. Cette méthode supervisée permet de prédire l'appartenance d'un individu à un cluster en fonction des variables actives.



— **Précision obtenue : 95,60%**

— **Variables les plus influentes (Golden Questions) :**

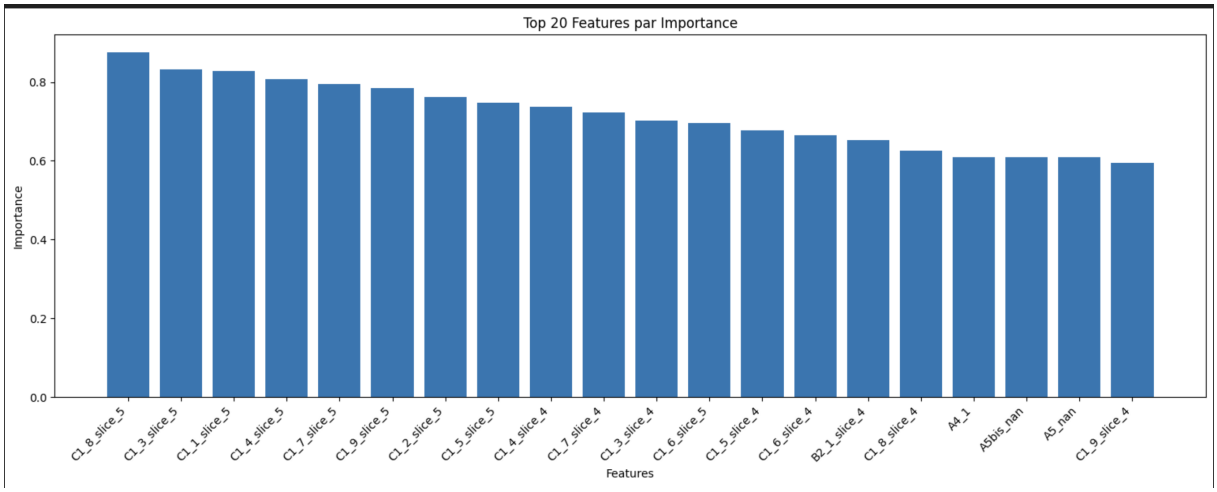
— A9_11

— A11_12

— A10_7

4.1.2 vert

Nous avons fait de meme pour les variables vertes



— **Précision obtenue : 96%**

— **Variables les plus influentes (Golden Questions) :**

— C1_8

— C1_3

— C1_1

4.2 Utilisation des variables illustratives

L'affectation a été testée avec des variables illustratives, permettant d'évaluer la robustesse du modèle avec un nombre réduit de variables.

Scénario	**Précision**
ORANGE vers VERT	51,9%
VERT vers ORANGE	70,3%

TABLE 2 – Précision de la réaffectation avec les variables illustratives

4.3 Matrices de confusion

Les matrices de confusion suivantes illustrent les performances de la réaffectation entre les clusters ORANGE et VERT.

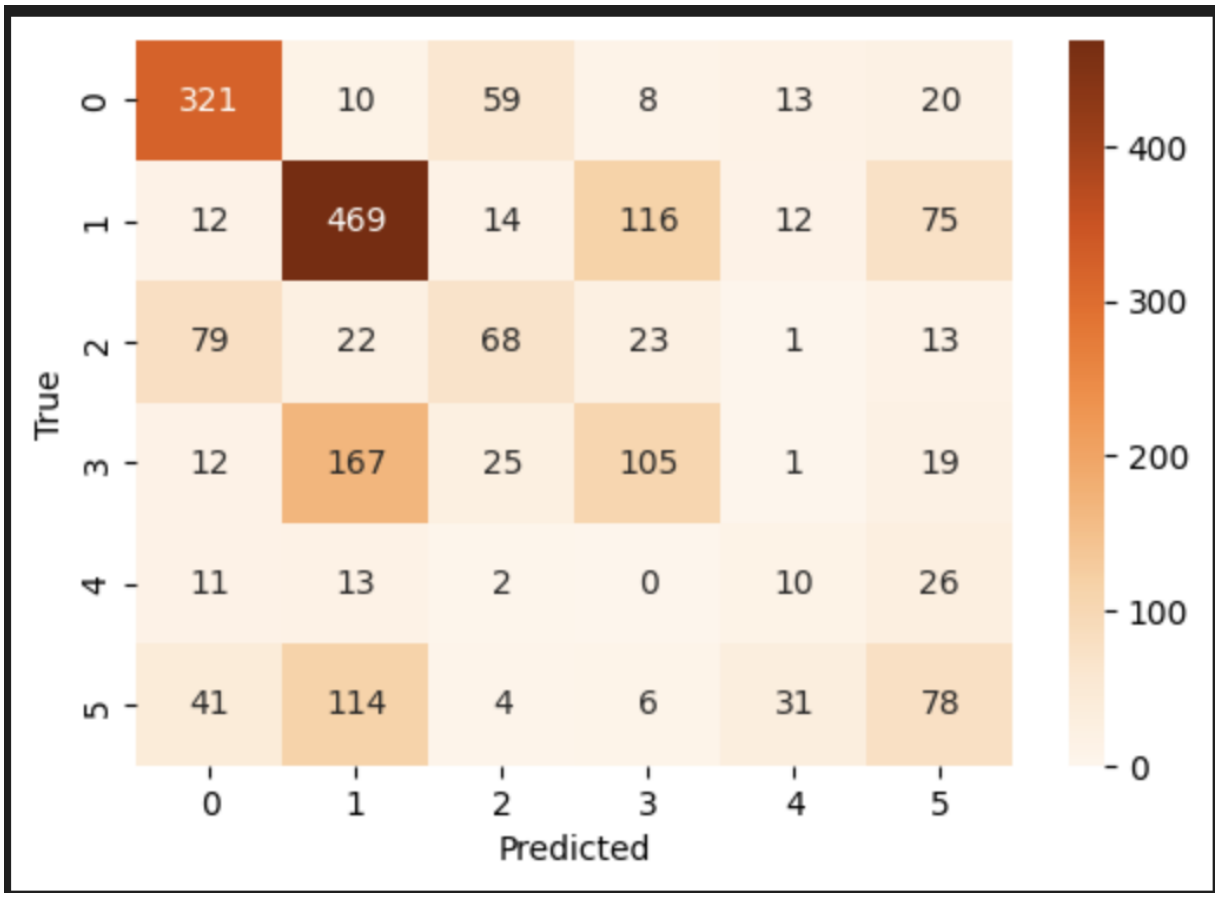


FIGURE 3 – Matrice de confusion : ORANGE vers VERT

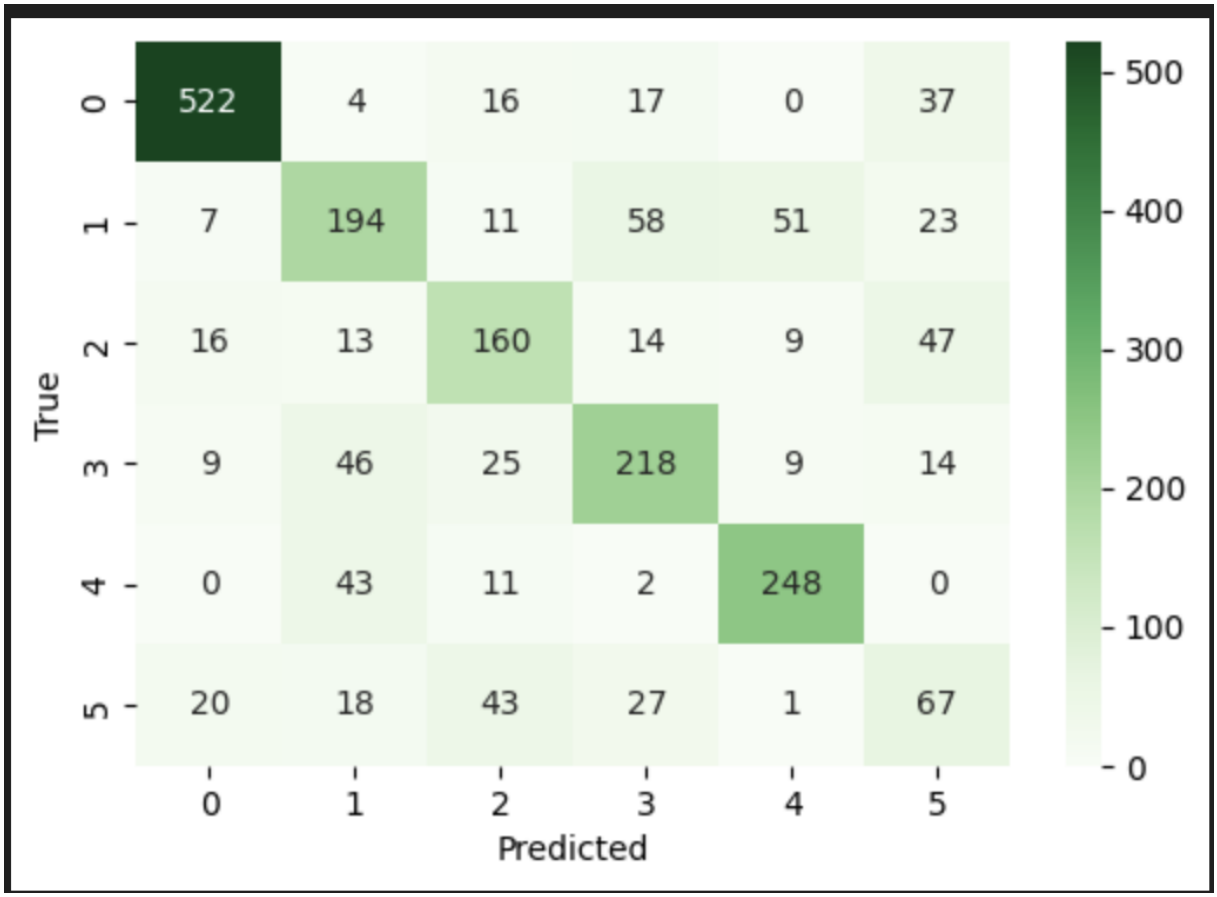


FIGURE 4 – Matrice de confusion : VERT vers ORANGE

---

5 Analyse des résultats

— Le *clustering* K-Means s'est avéré le plus performant parmi les méthodes testées, comme le montrent les scores de silhouette supérieurs.

— La réaffectation avec les variables actives a montré une précision élevée (**95,60%**), indiquant une forte capacité prédictive du modèle.

— Les variables illustratives ont montré une performance inférieure, avec des précisions de 51,9% et 70,3%, suggérant des axes d'amélioration potentiels pour la sélection des variables.

---

6 Conclusion

Dans ce notebook, nous avons réalisé deux segmentations : la segmentation Orange (A9, A10, A11) et la segmentation Vert (A11, A12, A13, A14, A4, A5, A5bis, etc.), en justifiant le nombre optimal de clusters (par exemple, six) à l'aide du coefficient de silhouette et de la répartition des individus.

La réaffectation des individus sur un échantillon test, en utilisant les variables actives et la Régression Logistique, a montré une précision souvent supérieure à 90

Enfin, l'utilisation de variables illustratives (non incluses dans la segmentation initiale) a permis d'évaluer leur capacité explicative. Les résultats obtenus (50 à 65

Ces analyses nous permettent d'évaluer la robustesse des segmentations et d'identifier les variables les plus pertinentes pour une application future.