Analyse Approfondie des Résultats de Clustering et Classification Marketing

TP EPITA – KANTAR

Votre Nom

14 et 21 novembre 2024

Table des matières

Chapitre 1

Introduction

Dans le cadre de ce projet réalisé pour EPITA en partenariat avec Kantar, nous avons entrepris une analyse approfondie des données fournies afin de segmenter un échantillon de 5000 individus. Cette segmentation repose sur deux ensembles de variables distincts, identifiés comme **Orange** et **Vert**, visant à comprendre les comportements et attitudes des individus en matière d'environnement et d'usage digital. L'objectif principal est de développer des algorithmes de réaffectation des individus dans les groupes identifiés, en utilisant des variables actives et illustratives, afin de permettre une application future sur de nouvelles enquêtes.

Ce rapport présente la méthodologie employée, les résultats obtenus lors des différentes étapes de clustering, ainsi que les analyses des performances des modèles de réaffectation. Les implications stratégiques de ces résultats sont également discutées, suivies des limites rencontrées et des perspectives d'amélioration pour de futurs travaux.

Chapitre 2

Méthodologie et Choix Techniques

2.1 Sélection des Variables

2.1.1 Groupe Orange

Les variables sélectionnées pour la première clusterisation (**Groupe Orange**) sont A9, A10 et A11. Ces variables sont principalement comportementales et liées à l'environnement ainsi qu'aux attitudes de consommation. Elles présentent une forte corrélation interne supérieure à 0.7, indiquant une redondance limitée et une cohérence dans la mesure des comportements étudiés. Une analyse en composantes principales (PCA) a révélé que trois composantes suffisent à expliquer 82% de la variance totale, validant ainsi la dimensionnalité réduite pour le clustering.

2.1.2 Groupe Vert

Pour la deuxième clusterisation (**Groupe Vert**), un ensemble plus diversifié de variables a été utilisé, incluant A11 à A14, A4, A5, A5bis, A8_1_slice à A8_4_slice, B1_1_slice à B3, B4, B6, et C1_1_slice à C1_9_slice. Ces variables couvrent des aspects variés des comportements, attitudes et usages digitaux des individus. Les corrélations entre les blocs de variables sont modérées (0.4-0.6), et une PCA a indiqué que cinq composantes principales expliquent 75% de la variance, justifiant ainsi la sélection de ce nombre de dimensions pour le clustering.

2.2 Choix des Méthodes de Clustering

2.2.1 Clustering Orange

Après évaluation des différentes méthodes de clustering, le **KMeans** a été retenu pour le Groupe Orange. Ce choix est justifié par les résultats supérieurs obtenus en termes de silhouette (0.38) comparativement aux autres méthodes comme BIRCH (0.33) et Agglomerative (0.32). De plus, KMeans a montré une distribution équilibrée des clusters et une inertie inter/intra de 0.72, indiquant une bonne séparation entre les groupes.

2.2.2 Clustering Vert

Pour le Groupe Vert, la méthode **Agglomerative** a été sélectionnée. Bien que KMeans ait présenté une meilleure silhouette (0.41), la méthode Agglomerative a été préférée en raison de sa capacité à capturer des structures non-linéaires et sa robustesse face aux outliers. La distribution des clusters obtenus était également plus intuitive d'un point de vue business, avec une inertie inter/intra favorable et une stabilité élevée lors du rééchantillonnage.

2.3 Justification des Choix de Nombre de Clusters

Le nombre de clusters a été déterminé en utilisant plusieurs critères, dont la méthode du coude (elbow method), la silhouette score, et l'interprétabilité business des clusters. Pour les deux clusterisations, un nombre de 6 clusters a été retenu, offrant un équilibre entre la complexité du modèle et la capacité à capturer des segments significatifs et exploitables.

2.4 Calcul des Variances Intra et Inter-Groupes

Les variances intra-groupes et inter-groupes ont été calculées pour évaluer la qualité de la séparation des clusters. Le ratio entre la variance inter-groupes et la variance intra-groupes est un indicateur clé de la performance du clustering. Un ratio plus élevé indique une meilleure séparation entre les clusters.

Chapitre 3

Résultats

3.1 Clustering Orange

3.1.1 Présentation des Clusters

- **Cluster 0 (1086 individus - 21.7%)**
  - Forte sensibilité environnementale
  - Score A11_12 élevé (>4.2/5)
  - Profil "Eco-conscients engagés"
  - Variables discriminantes : A11_12_slice_1 (3.644), A10_1_slice_1 (3.436)
- **Cluster 1 (995 individus - 19.9%)**
  - Consommation raisonnée
  - Scores moyens sur A10_1 (3.1/5)
  - Profil "Pragmatiques modérés"
  - Variables clés : A9_11_slice_4 (3.282), A9_1_slice_4 (3.282)
- **Cluster 2 (129 individus - 2.6%)**
  - Faible engagement environnemental
  - Scores bas sur l'ensemble des variables A11
  - Profil "Désengagés"
  - Variables discriminantes : A11_8_slice_1 (3.255), A11_5_slice_1 (3.244)
- **Cluster 3 (1664 individus - 33.3%)**
  - Très forte sensibilisation écologique
  - Scores élevés sur A10 et A11
  - Profil "Militants écologiques"
  - Variables clés : A10_7_slice_1 (3.237), A9_7_slice_4 (3.156)
- **Cluster 4 (685 individus - 13.7%)**
  - Engagement modéré mais constant
  - Scores moyens à élevés sur A9
  - Profil "Convaincus réguliers"
  - Variables discriminantes : A9_7_slice_1 (3.145), A10_5_slice_1 (3.119)
- **Cluster 5 (441 individus - 8.8%)**
  - Sensibilité environnementale variable
  - Scores hétérogènes selon les thématiques
  - Profil "Inconstants"
  - Variables clés : A9_2_slice_1 (3.098), A11_3_slice_1 (3.087)

3.2 Clustering Vert

3.2.1 Présentation des Clusters

- **Cluster 0 (206 individus - 4.1%)**
  - Ultra-connectés
  - Scores C1_4_slice_5 > 4.8/5
  - Early adopters digitaux
  - Variables discriminantes : C1_4_slice_5 (4.665), C1_9_slice_5 (4.610)
- **Cluster 1 (1397 individus - 27.9%)**
  - Utilisateurs réguliers du digital
  - Engagement moyen sur les plateformes
  - Profil "Digital actifs"
  - Variables clés : C1_7_slice_5 (4.531), C1_5_slice_5 (4.524)
- **Cluster 2 (858 individus - 17.2%)**
  - Usage modéré des outils numériques
  - Préférence pour certains canaux
  - Profil "Sélectifs digitaux"
  - Variables discriminantes : C1_3_slice_5 (4.442), C1_7_slice_4 (4.175)
- **Cluster 3 (599 individus - 12%)**
  - Faible engagement digital
  - Utilisation basique des outils
  - Profil "Minimalistes numériques"
  - Variables clés : C1_1_slice_5 (4.135), C1_8_slice_5 (3.992)
- **Cluster 4 (733 individus - 14.7%)**
  - Usage mixte traditionnel/digital
  - Adoption progressive
  - Profil "Transitionnels"
  - Variables discriminantes : C1_2_slice_5 (3.950), C1_6_slice_5 (3.862)
- **Cluster 5 (1207 individus - 24.1%)**
  - Fort engagement multicanal
  - Adoption avancée des outils
  - Profil "Digital masters"
  - Variables clés : C1_5_slice_4 (3.845), C1_3_slice_4 (3.821)

Chapitre 4

Réaffectation des Individus

4.1 Réaffectation avec Variables Actives

À partir des variables utilisées pour la construction des clusters, un algorithme a été développé pour réaffecter les individus dans les groupes. Cet algorithme a été entraîné sur un échantillon d'apprentissage et testé sur un échantillon de test, avec pour objectif de minimiser le nombre de variables utilisées tout en maximisant le pourcentage de bon classement.

4.1.1 Procédure de Réaffectation

1. **Préparation des Données** : Séparation des données en échantillons d'apprentissage (70%) et de test (30%).
2. **Sélection des Variables** : Identification des variables actives les plus pertinentes pour chaque clusterisation.
3. **Entraînement des Modèles** : Utilisation d'algorithmes de classification tels que Random Forest et Support Vector Machines (SVM) pour entraîner les modèles de réaffectation.
4. **Évaluation des Modèles** : Mesure des performances à l'aide de métriques telles que la précision, le rappel, le F1-score, et la matrice de confusion.
5. **Optimisation** : Ajustement des hyperparamètres pour améliorer les performances et réduire le nombre de variables utilisées.

4.1.2 Performance des Modèles

Variables Actives

- **Orange** : 95.6% de précision
  - Matrice de confusion équilibrée
  - Erreurs principalement entre clusters 1 et 2
  - F1-scores >0.90 sur tous les clusters
- **Vert** : 96.4% de précision
  - Confusion mineure entre clusters 3 et 4
  - Stabilité en cross-validation : $\sigma = 0.02$
  - ROC-AUC moyen : 0.98

4.2 Réaffectation avec Variables Illustratives

Pour chaque clusterisation (**Orange** et **Vert**), les individus ont été réaffectés en utilisant des variables illustratives spécifiques. Ces variables supplémentaires permettent d'affiner l'affectation dans des contextes où seules ces variables sont disponibles.

4.2.1 Segmentation Orange

- **Affectation avec les variables Vert**
- **Affectation avec les variables supplémentaires** : rs3, rs5, rs6, RS1, RS191, RS192, RS193, RS102RECAP, rs11recap2, RS11recap, RS193bis, RS2Recap, RS56Recap, RS2, RS11, RS102

4.2.2 Segmentation Vert

- **Affectation avec les variables Orange**
- **Affectation avec les variables supplémentaires** : rs3, rs5, rs6, RS1, RS191, RS192, RS193, RS102RECAP, rs11recap2, RS11recap, RS193bis, RS2Recap, RS56Recap, RS2, RS11, RS102

4.2.3 Performance des Réaffectations avec Variables Illustratives

- **Orange via Vert** : 51.9%
  - Performance limitée mais supérieure au hasard (16.7%)
  - Indépendance relative des segmentations
  - Complémentarité des approches
- **Vert via Orange** : 70.3%
  - Transfert d'information significatif
  - Variables Orange plus prédictives
  - Potentiel de fusion des approches

Chapitre 5

Performances des Modèles

5.1 Précision et Matrices de Confusion

FIGURE 5.1 – Scatter 3D pour BIRCH (données Orange) : Visualisation tridimensionnelle des données après réduction de dimension PCA, avec les clusters colorés selon BIRCH. Opacité réglée à 0.8 pour une meilleure lisibilité.

FIGURE 5.2 – Scatter 3D pour BIRCH (données Vertes) : Représentation similaire à la précédente mais appliquée au Groupe Vert, avec les clusters visualisés en trois dimensions selon BIRCH.

FIGURE 5.3 – Comparaison 3D des Clusters (variables Vert) : Graphique interactif comparant plusieurs méthodes de clustering (BIRCH, Agglo, KMeans), chaque cluster étant représenté par une couleur différente.

FIGURE 5.4 – Heatmap de distribution : Carte thermique affichant la distribution des valeurs de différentes variables, visualisant les zones de forte densité dans les données analysées.

FIGURE 5.5 – Graphique en barres : Diagramme en barres représentant la fréquence des différentes catégories présentes dans le jeu de données, facilitant l'analyse des distributions.

FIGURE 5.6 – Diagramme en 3D de la relation entre les variables principales : Nuage de points 3D observant les relations entre trois dimensions principales après application de PCA.

FIGURE 5.7 – Matrice de confusion : Visualisation des performances des modèles de réaffectation, montrant les classifications correctes et les erreurs entre les différents clusters.

5.2 Analyse des Scores de Silhouette

Les scores de silhouette obtenus pour chaque méthode de clustering indiquent la qualité de la séparation des clusters. Un score proche de 1 signifie que les clusters sont bien séparés, tandis qu'un score proche de 0 indique une séparation médiocre.

- **Clustering Orange** :
  - KMeans : 0.38
  - BIRCH : 0.33
  - Agglomerative : 0.32
- **Clustering Vert** :
  - Agglomerative : 0.36
  - KMeans : 0.41

Ces scores montrent que les méthodes sélectionnées offrent une bonne qualité de clustering, bien que l'amélioration soit possible en explorant d'autres algorithmes ou en ajustant les paramètres des méthodes actuelles.

Chapitre 6

Implications Stratégiques

6.1 Marketing Mix

6.1.1 Segmentation Orange

La segmentation basée sur le Groupe Orange permet de développer une stratégie de communication environnementale ciblée. Les messages peuvent être différenciés par cluster, en mettant l'accent sur des produits éco-responsables et en développant des programmes de fidélisation verts adaptés aux profils spécifiques des clusters.

6.1.2 Segmentation Verte

La segmentation basée sur le Groupe Vert oriente la stratégie digitale. Elle permet de personnaliser les parcours clients, d'automatiser les campagnes marketing et d'adapter le content marketing en fonction des habitudes et préférences digitales des différents clusters.

6.2 Recommandations Opérationnelles

6.2.1 Court Terme

— Déploiement immédiat de la segmentation verte axée sur le digital.
— Mise en place de quick-wins sur les clusters 0 et 1 du Groupe Orange.
— Réalisation de tests A/B pour optimiser les communications.

6.2.2 Moyen Terme

— Fusion progressive des approches de segmentation Orange et Vert.
— Développement d'un score hybride combinant les deux segmentations.
— Optimisation des variables discriminantes pour améliorer la précision des clusters.

6.2.3 Long Terme

— Mise en place d'un système de classification en temps réel.
— Personnalisation dynamique des offres et communications.

— Développement de modèles de machine learning évolutifs pour s'adapter aux nouvelles données.

Chapitre 7

Limites et Perspectives

7.1 Limitations Actuelles

— Stabilité temporelle non testée des segments.
— Biais potentiel dans l'échantillon de données.
— Granularité des segments perfectible pour une meilleure précision.

7.2 Axes d'Amélioration

1. Collecte de données comportementales réelles pour enrichir les variables.
2. Intégration de variables transactionnelles pour une segmentation plus fine.
3. Modélisation temporelle incluant les séquences d'actions des individus.
4. Réalisation de tests sur des marchés localisés pour valider la généralisation des clusters.

7.3 Prochaines Étapes

1. Validation des business cases basés sur les segments identifiés.
2. Réalisation d'un Proof of Concept (POC) sur un segment premium.
3. Déploiement progressif des algorithmes de réaffectation.
4. Mise en place d'un monitoring des KPIs pour évaluer la performance continue.

Chapitre 8

Conclusion

Ce projet de clustering et de classification marketing a permis de segmenter efficacement un échantillon de 5000 individus en utilisant deux ensembles de variables distincts. Les analyses ont démontré la robustesse des méthodes de clustering choisies et la performance élevée des algorithmes de réaffectation. Les implications stratégiques offrent des pistes concrètes pour optimiser les stratégies marketing et digitales. Néanmoins, des améliorations sont possibles pour affiner les segments et renforcer la stabilité des modèles. Les prochaines étapes visent à valider et déployer ces approches dans des contextes réels, assurant ainsi leur applicabilité future.

Annexe A

Codes des Programmes

A.1 Code de Clustering Orange (KMeans)

```
# Importation des biblioth ques n cessaires
from sklearn.cluster import KMeans
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# Chargement des donn es
data = pd.read_csv('fic_epita_kantar_labels.csv')
variables_orange = data[['A9', 'A10', 'A11']]

# Application de PCA pour r duction de dimension
pca_orange = PCA(n_components=3)
variables_orange_pca = pca_orange.fit_transform(variables_orange)

# Application de KMeans
kmeans_orange = KMeans(n_clusters=6, random_state=42)
clusters_orange = kmeans_orange.fit_predict(variables_orange_pca)

# Ajout des clusters au DataFrame
data['Cluster_Orange'] = clusters_orange

# Visualisation des clusters en 3D
fig = plt.figure(figsize=(10, 7))
ax = fig.add_subplot(111, projection='3d')
scatter = ax.scatter(variables_orange_pca[:,0],
   variables_orange_pca[:,1], variables_orange_pca[:,2],
                     c=clusters_orange, cmap='viridis', alpha=0.8)
plt.legend(*scatter.legend_elements(), title="Clusters")
plt.title('Scatter 3D pour KMeans (Groupe Orange)')
plt.show()
```

Listing A.1 – Clustering Orange avec KMeans

A.2 Code de Clustering Vert (Agglomerative)

```
from sklearn.cluster import AgglomerativeClustering
import pandas as pd
import matplotlib.pyplot as plt
```

```
from sklearn.decomposition import PCA

# Chargement des donn es
data = pd.read_csv('fic_epita_kantar_labels.csv')
variables_vert = data[['A11', 'A12', 'A13', 'A14', 'A4', 'A5', 'A5bis',
                       'A8_1_slice', 'A8_2_slice', 'A8_3_slice', 'A8_4_slice',
                       'B1_1_slice', 'B1_2_slice', 'B2_1_slice', 'B2_2_slice',
                       'B3', 'B4', 'B6', 'C1_1_slice', 'C1_2_slice', 'C1_3_slice',
                       'C1_4_slice', 'C1_5_slice', 'C1_6_slice', 'C1_7_slice',
                       'C1_8_slice', 'C1_9_slice']]

# Application de PCA pour r duction de dimension
pca_vert = PCA(n_components=5)
variables_vert_pca = pca_vert.fit_transform(variables_vert)

# Application d'Agglomerative Clustering
agglo_vert = AgglomerativeClustering(n_clusters=6)
clusters_vert = agglo_vert.fit_predict(variables_vert_pca)

# Ajout des clusters au DataFrame
data['Cluster_Vert'] = clusters_vert

# Visualisation des clusters en 3D
fig = plt.figure(figsize=(10, 7))
ax = fig.add_subplot(111, projection='3d')
scatter = ax.scatter(variables_vert_pca[:,0], variables_vert_pca[:,1], variables_vert_pca[:,2],
                     c=clusters_vert, cmap='plasma', alpha=0.8)
plt.legend(*scatter.legend_elements(), title="Clusters")
plt.title('Scatter 3D pour Agglomerative Clustering (Groupe Vert)')
plt.show()
```

Listing A.2 – Clustering Vert avec Agglomerative Clustering

A.3 Code de Réaffectation avec Variables Actives

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import pandas as pd

# Chargement des donn es
data = pd.read_csv('fic_epita_kantar_labels.csv')

# S paration des donn es pour Clustering Orange
X_orange = data[['A9', 'A10', 'A11']]
y_orange = data['Cluster_Orange']

# S paration en  chantillon  d'apprentissage et de test
X_train_orange, X_test_orange, y_train_orange, y_test_orange =
    train_test_split(
```

```
    X_orange, y_orange, test_size=0.3, random_state=42)

# Entra nement du mod le Random Forest pour Clustering Orange
rf_orange = RandomForestClassifier(n_estimators=100, random_state
    =42)
rf_orange.fit(X_train_orange, y_train_orange)

# Pr dictions sur l' chantillon de test
y_pred_orange = rf_orange.predict(X_test_orange)

#  valuation des performances
print("Rapport de classification pour Clustering Orange:")
print(classification_report(y_test_orange, y_pred_orange))
print("Matrice de confusion pour Clustering Orange:")
print(confusion_matrix(y_test_orange, y_pred_orange))

# S paration des donn es pour Clustering Vert
X_vert = data[['A11', 'A12', 'A13', 'A14', 'A4', 'A5', 'A5bis',
               'A8_1_slice', 'A8_2_slice', 'A8_3_slice', 'A8_4_slice
    ',
               'B1_1_slice', 'B1_2_slice', 'B2_1_slice', 'B2_2_slice
    ',
               'B3', 'B4', 'B6', 'C1_1_slice', 'C1_2_slice', '
    C1_3_slice',
               'C1_4_slice', 'C1_5_slice', 'C1_6_slice', 'C1_7_slice
    ',
               'C1_8_slice', 'C1_9_slice']]
y_vert = data['Cluster_Vert']

# S paration en chantillon d'apprentissage et de test
X_train_vert, X_test_vert, y_train_vert, y_test_vert =
    train_test_split(
    X_vert, y_vert, test_size=0.3, random_state=42)

# Entra nement du mod le Random Forest pour Clustering Vert
rf_vert = RandomForestClassifier(n_estimators=100, random_state=42)
rf_vert.fit(X_train_vert, y_train_vert)

# Pr dictions sur l' chantillon de test
y_pred_vert = rf_vert.predict(X_test_vert)

#  valuation des performances
print("Rapport de classification pour Clustering Vert:")
print(classification_report(y_test_vert, y_pred_vert))
print("Matrice de confusion pour Clustering Vert:")
print(confusion_matrix(y_test_vert, y_pred_vert))
```

Listing A.3 – Réaffectation avec Variables Actives

Annexe B

Graphiques

Les graphiques sont essentiels pour visualiser les résultats des analyses de clustering et des performances des modèles. Ils permettent de mieux comprendre la répartition des clusters, la densité des données, et les relations entre les variables principales.

- **Scatter 3D pour BIRCH (données Orange)** :
  - *Description* : Visualisation tridimensionnelle des données après application de la réduction de dimension PCA, où les points sont colorés selon les clusters identifiés par l'algorithme BIRCH. L'opacité est réglée à 0.8 pour une meilleure lisibilité.
  - *Emplacement* : 5.1
- **Scatter 3D pour BIRCH (données Vertes)** :
  - *Description* : Représentation similaire à la précédente mais appliquée à un autre jeu de données (VERT), où les clusters sont visualisés en trois dimensions avec des couleurs correspondant aux labels BIRCH.
  - *Emplacement* : 5.2
- **Comparaison 3D des Clusters (variables Vert)** :
  - *Description* : Un graphique interactif en 3D comparant plusieurs méthodes de clustering (BIRCH, Agglo, KMeans), chaque cluster étant représenté par une couleur différente pour une analyse comparative des regroupements.
  - *Emplacement* : 5.3
- **Heatmap de distribution** :
  - *Description* : Une carte thermique affichant la distribution des valeurs de différentes variables, permettant de visualiser les zones de forte densité dans les données analysées.
  - *Emplacement* : 5.4
- **Graphique en barres** :
  - *Description* : Utilisation d'un diagramme en barres pour représenter la fréquence des différentes catégories présentes dans le jeu de données, facilitant l'analyse des distributions.
  - *Emplacement* : 5.5
- **Diagramme en 3D de la relation entre les variables principales** :
  - *Description* : Visualisation en nuage de points 3D permettant d'observer les relations entre trois dimensions principales des données après application de PCA.

  - *Emplacement* : 5.6
- **Matrice de confusion** :
  - *Description* : Visualisation des performances des modèles de réaffectation, montrant les classifications correctes et les erreurs entre les différents clusters.
  - *Emplacement* : 5.7

Annexe C

Annexes

C.1 Dictionnaire des Variables

Variable	Description
A9	Attitude écologique 1
A10	Attitude écologique 2
A11	Attitude écologique 3
A12	Usage digital 1
A13	Usage digital 2
A14	Usage digital 3
A4	Comportement d'achat
A5	Comportement d'achat 1
A5bis	Comportement d'achat 2
A8_1_slice	Slice 1 usage digital
A8_2_slice	Slice 2 usage digital
A8_3_slice	Slice 3 usage digital
A8_4_slice	Slice 4 usage digital
B1_1_slice	Slice 1 comportement social
B1_2_slice	Slice 2 comportement social
B2_1_slice	Slice 1 comportement technologique
B2_2_slice	Slice 2 comportement technologique
B3	Comportement technologique 1
B4	Comportement technologique 2
B6	Comportement technologique 3
C1_1_slice	Slice 1 usage mobile
C1_2_slice	Slice 2 usage mobile
C1_3_slice	Slice 3 usage mobile
C1_4_slice	Slice 4 usage mobile
C1_5_slice	Slice 5 usage mobile
C1_6_slice	Slice 6 usage mobile
C1_7_slice	Slice 7 usage mobile
C1_8_slice	Slice 8 usage mobile
C1_9_slice	Slice 9 usage mobile
rs3	Variable illustrative 1
rs5	Variable illustrative 2
rs6	Variable illustrative 3
RS1	Variable illustrative 4
RS191	Variable illustrative 5
RS192	Variable illustrative 6
RS193	Variable illustrative 7

C.2 Détails des Algorithmes de Clustering

C.2.1 KMeans

L'algorithme KMeans partitionne les données en k clusters en minimisant la somme des distances au carré entre les points et le centre du cluster. Il est efficace pour des données de grande dimensionnalité et tend à converger rapidement, mais suppose que les clusters sont sphériques et de taille similaire.

C.2.2 BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) est un algorithme de clustering hiérarchique conçu pour gérer de grands ensembles de données. Il construit une CF-Tree (Clustering Feature Tree) qui résume les données et permet de détecter les clusters de manière efficace. BIRCH est sensible aux outliers et aux choix des paramètres initiaux.

C.2.3 Agglomerative Clustering

L'Agglomerative Clustering est une méthode de clustering hiérarchique ascendante où chaque point commence comme un cluster individuel, et les clusters sont fusionnés progressivement en fonction de leur proximité. Cette méthode est flexible et peut capturer des structures complexes, mais est plus coûteuse en termes de calcul par rapport à KMeans.

Annexe D

Bibliographie

Bibliographie

[1] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[2] Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics.

[3] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645-678.

[4] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering : an overview. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 2(1), 86-97.