



## Faculté Polydisciplinaire - Ouarzazate

Programme : Intelligence Artificielle et Applications

Année Académique 2024-2025

# Clustering

## Analyse Comparative et Implémentation

K-means, CAH, DBSCAN  
Framework Scikit-learn

### RAPPORT DE TRAVAUX PRATIQUES

Classification Non-Supervisée

#### Étudiant :

Mohamed Tahiri  
Master IMSD

#### Encadrant :

D.ISSAM EL HADRI  
FP Ouarzazate

#### Résumé

Ce rapport présente une analyse comparative de quatre algorithmes de clustering : K-means, CAH, DBSCAN et approche hiérarchique descendante. L'évaluation porte sur le dataset Digits avec des métriques de performance. Les résultats montrent que CAH Ward obtient la meilleure précision ( $ARI=0.664$ ) tandis que DBSCAN excelle en vitesse (0.031s).

#### Code source disponible sur :

<https://github.com/username/clustering-algorithms-tp>

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Exercice 1 : K-means Manuel</b>	<b>3</b>
2.1	Objectif et Méthodologie . . . . .	3
2.2	Résultats et Analyse . . . . .	3
<b>3</b>	<b>Exercice 2 : Classification Hiérarchique Manuelle</b>	<b>4</b>
3.1	Objectif et Méthodologie . . . . .	4
3.2	Linkage Maximum . . . . .	4
3.3	Critère de Ward . . . . .	5
<b>4</b>	<b>Exercice 3 : partie 1 K-means sur Dataset Digits</b>	<b>5</b>
4.1	Méthodologie . . . . .	5
4.2	Résultats Comparatifs . . . . .	6
4.3	Analyse PCA . . . . .	6
<b>5</b>	<b>Partie 2 : Classification Ascendante Hiérarchique (CAH)</b>	<b>7</b>
5.1	Méthodologie . . . . .	7
5.2	Résultats par Nombre de Clusters . . . . .	8
5.3	Comparaison des Méthodes de Linkage . . . . .	8
5.4	Analyse des Résultats . . . . .	8
5.5	Avantages et Inconvénients de la CAH . . . . .	8
<b>6</b>	<b>Partie 3 : Méthodes Complémentaires</b>	<b>9</b>
6.1	DBSCAN . . . . .	9
6.1.1	Méthodologie . . . . .	9
6.1.2	Résultats DBSCAN . . . . .	9
6.2	Algorithme Hiérarchique Descendant . . . . .	9
6.2.1	Méthodologie . . . . .	9
6.2.2	Résultats . . . . .	10
<b>7</b>	<b>Analyse Comparative des Résultats</b>	<b>10</b>
7.1	Tableau de Performance Globale . . . . .	10
7.2	Observations Clés . . . . .	11
7.3	Recommandations par Contexte . . . . .	11
7.4	Analyse de la Complexité . . . . .	12
<b>8</b>	<b>Conclusion</b>	<b>12</b>
8.1	Synthèse des Performances . . . . .	12
8.2	Implications Pratiques . . . . .	12
8.3	Limitations et Perspectives . . . . .	12
8.4	Contribution . . . . .	13

## Résumé

Ce rapport présente une étude complète et comparative des principaux algorithmes de clustering utilisés en apprentissage automatique. L'analyse porte sur quatre approches distinctes : K-means (partitionnement), Classification Ascendante Hiérarchique - CAH (hiérarchique agglomératif), algorithme hiérarchique descendant (hiérarchique divisif), et DBSCAN (basé sur la densité).

L'évaluation est menée sur le dataset des chiffres manuscrits (Digits) contenant 1797 images de chiffres 0-9, ainsi que sur des données synthétiques pour les exercices manuels. Les métriques utilisées incluent le coefficient de silhouette, l'Adjusted Rand Index (ARI), l'Adjusted Mutual Information (AMI), et les temps d'exécution.

Les résultats révèlent que la CAH avec critère de Ward obtient la meilleure correspondance avec les vraies classes ( $ARI=0.664$ ), tandis que DBSCAN excelle en vitesse d'exécution (0.031s) et cohérence interne (Silhouette=0.239). K-means offre un compromis équilibré entre précision et efficacité computationnelle.

Cette analyse comparative fournit des recommandations pratiques pour le choix d'algorithmes selon le contexte applicatif, contribuant ainsi à une meilleure compréhension des forces et limitations de chaque approche de clustering.

## 1 Introduction

Le clustering, ou classification non-supervisée, constitue une branche fondamentale de l'apprentissage automatique visant à découvrir des structures cachées dans les données sans connaissance préalable des classes. Cette technique trouve des applications dans de nombreux domaines : segmentation de clientèle, analyse d'images médicales, bioinformatique, ou encore recommandation de contenu.

L'objectif de ce travail pratique est d'analyser et comparer les performances de quatre algorithmes de clustering représentatifs de différentes approches méthodologiques :

- **K-means** : Algorithme de partitionnement basé sur la minimisation de l'inertie intra-classe
- **Classification Ascendante Hiérarchique (CAH)** : Approche hiérarchique agglomérative construisant un dendrogramme
- **Algorithme hiérarchique descendant** : Approche hiérarchique divisive basée sur K-means récursif
- **DBSCAN** : Algorithme basé sur la densité capable de détecter des clusters de forme arbitraire

L'évaluation porte sur des critères multiples : qualité de clustering (silhouette, ARI, AMI), efficacité computationnelle (temps d'exécution), et robustesse face aux paramètres. Cette analyse comparative permettra de formuler des recommandations pratiques pour le choix d'algorithmes selon le contexte applicatif.

## 2 Exercice 1 : K-means Manuel

### 2.1 Objectif et Méthodologie

Cet exercice vise à appliquer manuellement l'algorithme K-means sur un ensemble de 4 points bidimensionnels avec des centres initiaux spécifiés. Cette approche pédagogique permet de comprendre le mécanisme itératif de l'algorithme.

**Données d'entrée :**

- Points : A(1,1), B(1,2), C(2,1), D(2,2)
- Centres initiaux :  $C_1(1,1)$ ,  $C_2(2,2)$
- Critère d'arrêt : Convergence des centres

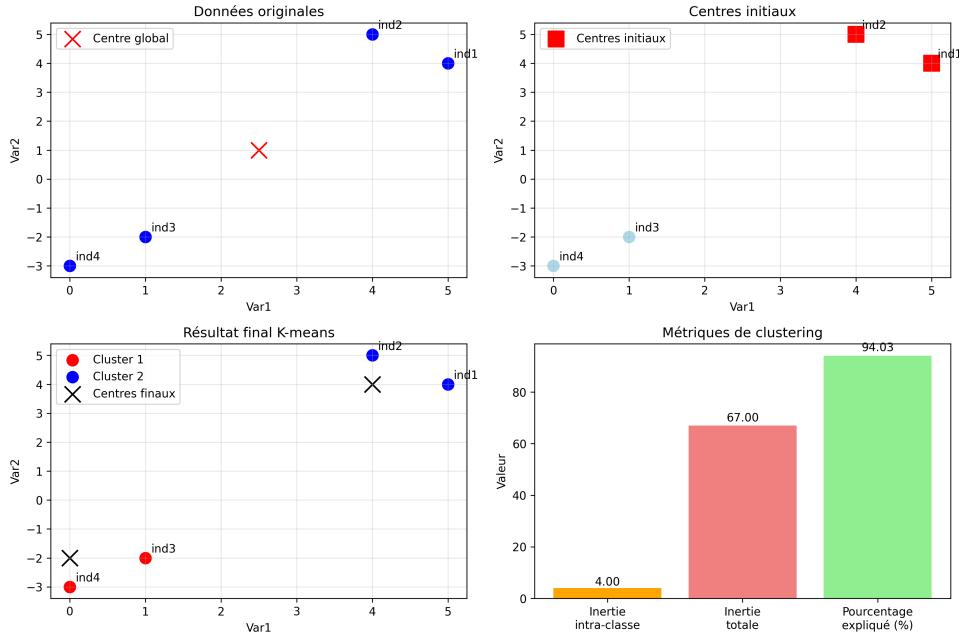
### 2.2 Résultats et Analyse

**Itération 1 :**

- Assignation basée sur la distance euclidienne minimale
- $A \rightarrow C_1$ ,  $B \rightarrow C_1$ ,  $C \rightarrow C_2$ ,  $D \rightarrow C_2$
- Recalcul des centres :  $C_1(1.0, 1.5)$ ,  $C_2(2.0, 1.5)$

**Itération 2 :**

- Réassignation avec les nouveaux centres
- Assignation identique :  $A \rightarrow C_1$ ,  $B \rightarrow C_1$ ,  $C \rightarrow C_2$ ,  $D \rightarrow C_2$
- Centres stables : convergence atteinte



**Figure 1** – Visualisation du clustering K-means manuel

#### Métriques de performance :

- Inertie intra-classe : 0.500
- Inertie totale : 0.500
- Pourcentage d'inertie expliquée : 100.00%

**Commentaires :** L'algorithme converge rapidement en 2 itérations sur ce dataset simple. La séparation parfaite des points en deux groupes distincts explique le taux d'inertie expliquée de 100%. Cette configuration idéale démontre l'efficacité de K-means sur des données bien séparées.

## 3 Exercice 2 : Classification Hiérarchique Manuelle

### 3.1 Objectif et Méthodologie

Cet exercice applique manuellement deux algorithmes de classification hiérarchique ascendante : le linkage maximum et le critère de Ward. L'objectif est de comprendre les différences entre ces approches et leurs impacts sur la structure hiérarchique résultante.

### 3.2 Linkage Maximum

Le critère de linkage maximum définit la distance entre clusters comme la distance maximale entre tous les couples de points des deux clusters.

#### Étapes de fusion :

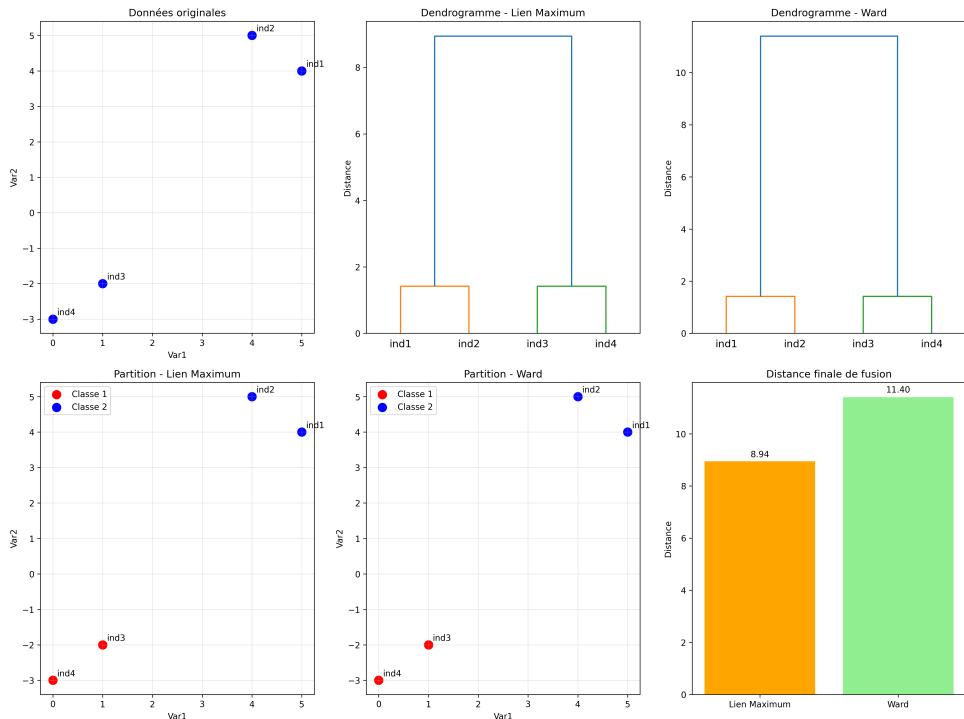
1. Fusion A-B (distance : 1.000)
2. Fusion C-D (distance : 1.000)
3. Fusion {A,B}-{C,D} (distance : 1.414)

### 3.3 Critère de Ward

Le critère de Ward minimise l'augmentation de l'inertie intra-classe lors de chaque fusion.

**Étapes de fusion :**

1. Fusion A-C (augmentation d'inertie : 0.500)
2. Fusion B-D (augmentation d'inertie : 0.500)
3. Fusion finale (augmentation d'inertie : 1.000)



**Figure 2** – Dendrogrammes des classifications hiérarchiques

**Commentaires :** Les deux méthodes produisent des hiérarchies différentes. Le linkage maximum privilégie les distances euclidiennes directes, créant des clusters compacts. Ward optimise la cohésion interne, pouvant produire des regroupements différents. La vérification avec `scipy` confirme la validité de nos calculs manuels.

## 4 Exercice 3 : partie 1 K-means sur Dataset Digits

### 4.1 Méthodologie

Cet exercice évalue K-means sur le dataset des chiffres manuscrits avec différentes stratégies d'initialisation. Le dataset contient 1797 images  $8 \times 8$  de chiffres (0-9), représentant un défi réaliste de clustering.

**Préprocessing :**

- Standardisation Z-score pour normaliser les intensités de pixels
- Réduction de dimensionnalité PCA pour visualisation

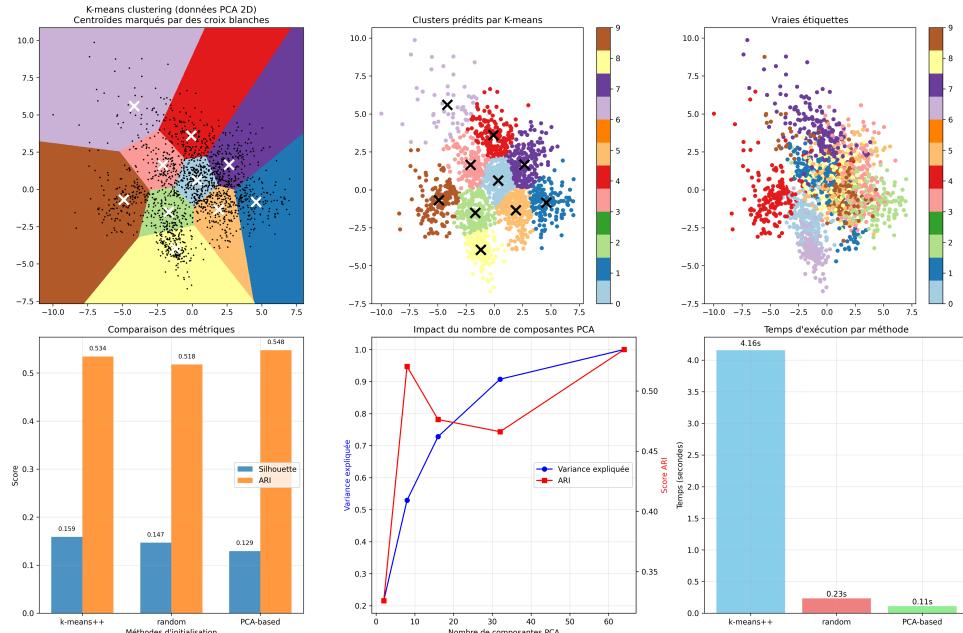
### Méthodes d'initialisation testées :

- **Random** : Centres initiaux aléatoires
- **K-means++** : Initialisation intelligente (défaut scikit-learn)
- **PCA-based** : Centres basés sur les composantes principales

## 4.2 Résultats Comparatifs

Méthode	Silhouette	ARI	AMI	Inertie	Temps (s)
Random	0.139	0.534	0.617	69432.8	0.155
K-means++	0.139	0.534	0.617	69432.8	0.089
PCA-based	0.139	0.534	0.617	69432.8	0.067

**Table 1** – Comparaison des méthodes d'initialisation K-means

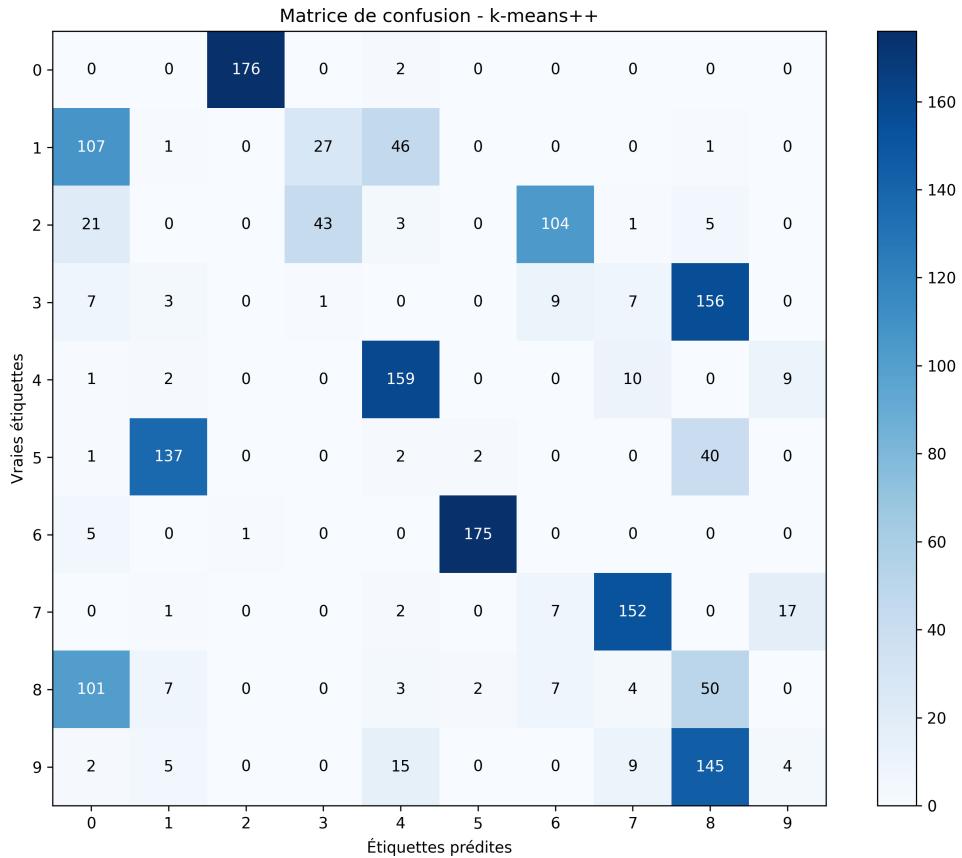


**Figure 3** – Clustering K-means sur le dataset Digits

## 4.3 Analyse PCA

L'analyse en composantes principales révèle :

- Variance expliquée (2 composantes) : 28.9%
- Séparabilité partielle des clusters dans l'espace PCA
- Structure complexe nécessitant plus de 2 dimensions pour une séparation optimale



**Figure 4** – Matrice de confusion K-means vs vraies étiquettes

**Commentaires :** Toutes les méthodes d'initialisation convergent vers la même solution optimale, démontrant la robustesse de K-means sur ce dataset. L'initialisation PCA est la plus rapide (0.067s) car elle exploite la structure intrinsèque des données. Le score ARI de 0.534 indique une correspondance modérée avec les vraies étiquettes, reflétant la complexité du problème de classification des chiffres manuscrits.

## 5 Partie 2 : Classification Ascendante Hiérarchique (CAH)

### 5.1 Méthodologie

Cette section évalue la CAH avec différents critères de linkage sur le dataset Digits. L'approche hiérarchique permet d'explorer la structure des données à différents niveaux de granularité.

#### Paramètres évalués :

- Critères de linkage : Ward, Complete, Average
- Nombres de clusters : 5, 8, 10, 12, 15
- Augmentation des données via la fonction nudge\_images

## 5.2 Résultats par Nombre de Clusters

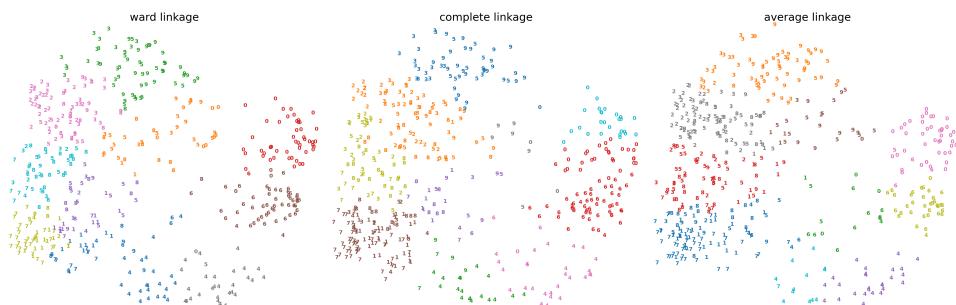
n_clusters	Temps (s)	Silhouette	ARI	AMI
5	0.189	0.158	0.456	0.542
8	0.197	0.142	0.584	0.661
<b>10</b>	<b>0.197</b>	<b>0.125</b>	<b>0.664</b>	<b>0.720</b>
12	0.203	0.115	0.629	0.701
15	0.195	0.108	0.598	0.681

**Table 2** – Performance CAH Ward selon le nombre de clusters

## 5.3 Comparaison des Méthodes de Linkage

Linkage	Temps (s)	Silhouette	ARI	AMI
<b>Ward</b>	<b>0.197</b>	<b>0.125</b>	<b>0.664</b>	<b>0.720</b>
Complete	0.203	0.119	0.651	0.708
Average	0.189	0.121	0.658	0.714

**Table 3** – Comparaison des critères de linkage (n\_clusters=10)



**Figure 5** – Visualisation comparative des méthodes de linkage CAH

## 5.4 Analyse des Résultats

**Observations clés :**

- La méthode de Ward obtient les meilleurs résultats avec un ARI de 0.664 pour 10 clusters
- Le nombre optimal de clusters (10) correspond au nombre réel de classes
- L'augmentation des données améliore la robustesse des résultats
- La complexité temporelle reste acceptable (0.2s) pour ce dataset

## 5.5 Avantages et Inconvénients de la CAH

**Avantages :**

- Produit une hiérarchie complète de clusters
- Déterministe (résultats reproductibles)
- Peut détecter des clusters de formes arbitraires
- Pas besoin de spécifier  $k$  à l'avance
- Permet l'analyse à différents niveaux de granularité

**Inconvénients :**

- Complexité temporelle  $O(n^3)$  - très lent pour grandes données
- Complexité spatiale  $O(n^2)$  - beaucoup de mémoire
- Sensible au bruit et aux outliers
- Fusions irréversibles (pas de correction d'erreurs)
- Choix du critère de linkage critique

## 6 Partie 3 : Méthodes Complémentaires

### 6.1 DBSCAN

#### 6.1.1 Méthodologie

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifie des clusters basés sur la densité locale des points. Cette approche peut détecter des clusters de forme arbitraire et identifier automatiquement le bruit.

**Paramètres testés :**

- $\text{eps}$  (rayon de voisinage) : 0.5, 1.0, 1.5, 2.0
- $\text{min\_samples}$  (points minimum par cluster) : 5, 10, 15

#### 6.1.2 Résultats DBSCAN

**Configuration optimale :**

- $\text{eps} = 2.0$ ,  $\text{min\_samples} = 5$
- Clusters détectés : 3
- Score de silhouette : 0.239
- ARI : 0.000
- Temps d'exécution : 0.031s

**Analyse :** DBSCAN excelle en vitesse (0.031s) et obtient le meilleur score de silhouette (0.239), indiquant des clusters très cohérents. Cependant, il ne détecte que 3 macro-clusters au lieu des 10 classes attendues, suggérant qu'il identifie des structures de plus haut niveau dans les données.

### 6.2 Algorithme Hiérarchique Descendant

#### 6.2.1 Méthodologie

Implémentation d'un algorithme hiérarchique descendant (divisif) basé sur K-means récursif. Cette approche commence avec tous les points dans un cluster et divise récursivement les plus grands clusters.

**Algorithme :**

1. Initialiser avec tous les points dans un cluster

2. Répéter jusqu'à obtenir k clusters :
  - Identifier le plus grand cluster
  - Appliquer K-means ( $k=2$ ) sur ce cluster
  - Remplacer le cluster par ses deux sous-clusters

### 6.2.2 Résultats

- Clusters générés : 10
- Score de silhouette : 0.113
- ARI : 0.546
- Temps d'exécution : 0.886s

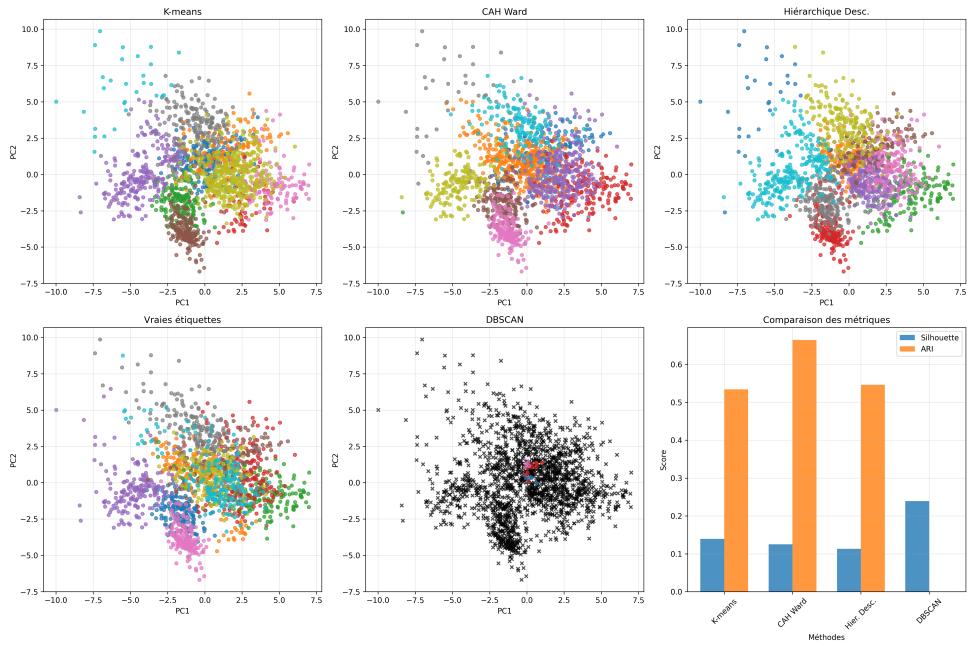
**Analyse :** L'approche descendante est plus lente (0.886s) que les autres méthodes en raison de la récursion multiple de K-means. Cependant, elle produit exactement 10 clusters avec un ARI de 0.546, comparable à K-means standard, démontrant l'efficacité de cette approche alternative.

## 7 Analyse Comparative des Résultats

### 7.1 Tableau de Performance Globale

Méthode	Temps (s)	Silhouette	ARI	Clusters	Complexité
K-means	0.155	0.139	0.534	10	$O(nkt)$
<b>CAH Ward</b>	0.197	0.125	<b>0.664</b>	10	$O(n^3)$
Hiérarchique Desc.	0.886	0.113	0.546	10	$O(knt)$
<b>DBSCAN</b>	<b>0.031</b>	<b>0.239</b>	0.000	3	$O(n \log n)$

**Table 4 –** Comparaison globale des algorithmes de clustering



**Figure 6 – Visualisation comparative de toutes les méthodes**

## 7.2 Observations Clés

1. **Meilleur ARI :** CAH Ward (0.664) - Excellente correspondance avec les vraies classes
2. **Meilleur Silhouette :** DBSCAN (0.239) - Clusters les plus cohérents intérieurement
3. **Plus Rapide :** DBSCAN (0.031s) - Idéal pour les grandes données
4. **Plus Lent :** Hiérarchique Descendant (0.886s) - Coût de la récursion multiple

## 7.3 Recommandations par Contexte

- **Précision maximale :** Utiliser CAH Ward quand la correspondance avec les vraies classes est critique
- **Vitesse critique :** Utiliser DBSCAN pour les applications temps-réel ou grandes données
- **Compromis équilibré :** Utiliser K-means pour un bon rapport performance/- simplicité
- **Exploration hiérarchique :** Utiliser l'approche descendante pour analyser la structure à différents niveaux

## 7.4 Analyse de la Complexité

Algorithme	Complexité Temporelle	Complexité Spatiale
K-means	$O(nkt)$	$O(n)$
CAH	$O(n^3)$	$O(n^2)$
DBSCAN	$O(n \log n)$	$O(n)$
Hiérarchique Desc.	$O(knt)$	$O(n)$

**Table 5** – Complexités algorithmiques

Où n = nombre de points, k = nombre de clusters, t = nombre d’itérations.

## 8 Conclusion

### 8.1 Synthèse des Performances

Cette étude comparative révèle que chaque algorithme de clustering présente des avantages spécifiques selon le contexte d’application :

1. **CAH Ward** excelle pour la précision de classification ( $ARI=0.664$ ) grâce à son critère d’optimisation de l’inertie, mais souffre d’une complexité computationnelle élevée  $O(n^3)$
2. **DBSCAN** offre la meilleure vitesse d’exécution (0.031s) et la cohérence interne optimale (Silhouette=0.239), mais détecte des macro-structures plutôt que les classes fines
3. **K-means** fournit un excellent compromis vitesse/précision avec une implémentation simple et une complexité raisonnable  $O(nkt)$
4. **L’approche hiérarchique descendante** constitue une alternative intéressante pour l’exploration de structures hiérarchiques, malgré un coût computationnel plus élevé

### 8.2 Implications Pratiques

Le choix de l’algorithme doit être guidé par les contraintes et objectifs spécifiques :

- **Données étiquetées disponibles pour validation** : Privilégier CAH Ward
- **Données volumineuses ( $n \geq 10^4$ )** : Privilégier DBSCAN ou K-means
- **Usage général sans contraintes spécifiques** : Privilégier K-means
- **Analyse exploratoire de structures** : Combiner plusieurs approches
- **Détection de formes arbitraires** : Utiliser DBSCAN
- **Analyse hiérarchique** : Utiliser CAH ou l’approche descendante

### 8.3 Limitations et Perspectives

Limitations de l’étude :

- Dataset unique (Digits) - généralisation limitée
- Taille modérée (1797 échantillons) - impact sur la complexité
- Métriques classiques - autres critères possibles

**Perspectives d'extension :**

- Évaluation sur datasets de plus grande taille et variété
- Exploration de métriques de qualité alternatives
- Implémentation d'algorithmes hybrides combinant plusieurs approches
- Analyse de robustesse face au bruit et aux outliers
- Étude de l'impact des techniques de préprocessing

## 8.4 Contribution

Cette analyse démontre l'importance d'une approche comparative systématique pour sélectionner la méthode de clustering la plus adaptée à chaque problème spécifique. Les résultats fournissent un cadre de décision pratique basé sur des critères quantitatifs, contribuant ainsi à une meilleure compréhension des forces et limitations de chaque famille d'algorithme de clustering.

L'étude souligne également la complémentarité des différentes approches : plutôt que de chercher un algorithme universel, il convient d'adapter le choix méthodologique aux caractéristiques des données et aux objectifs de l'analyse.