

# Analyse de Classification et Clustering de Vins

Étude Comparative des Algorithmes KNN, K-Means et CAH

Cours : Data Mining

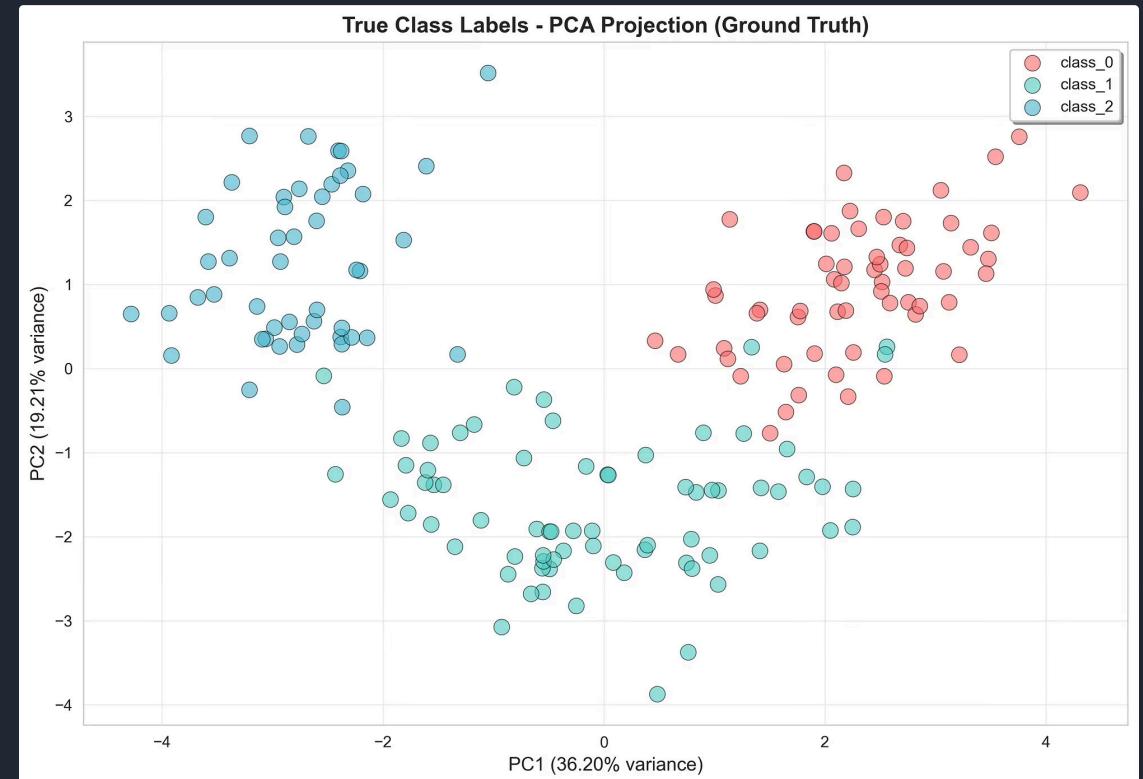
Date : 19/10/2025

Étudiants : Mathew Kristoffer Ewan KAPOOR - Raphael MARTIN - Julien KLINGER

# Introduction au Dataset

## Présentation du Jeu de Données

- Dataset Wine Recognition de sklearn
- 178 échantillons de vins issus de 3 cépages italiens
- 13 caractéristiques chimiques (alcool, acidité, phénols, etc.)
- Classes bien séparées, idéales pour démontrer les algorithmes



Légende : "Vérité terrain : Trois cépages montrent une séparation claire dans l'espace chimique"

# Aperçu Méthodologique

Trois Algorithmes, Trois Approches

## KNN (K-Plus Proches Voisins)

Classification Supervisée

- "Utilise des données d'entraînement étiquetées pour classifier de nouveaux vins"
- "Principe : un vin est similaire à ses K plus proches voisins"
- "Nous trouvons le K optimal via validation croisée"

## K-Means

Clustering Non Supervisé (Partitionnel)

- "Découvre des groupes sans connaître les vraies étiquettes"
- "Optimise itérativement les centroïdes des clusters"
- "Nécessite de spécifier K à l'avance - nous utilisons la méthode du coude"

## CAH (Clustering Hiérarchique)

Clustering Non Supervisé (Hiérarchique)

- "Construit un arbre de relations entre les échantillons"
- "Ne nécessite pas K à l'avance - nous coupons le dendrogramme"
- "Montre la hiérarchie complète, pas seulement le regroupement final"



Acquisition

Prétraitement

Algorithm

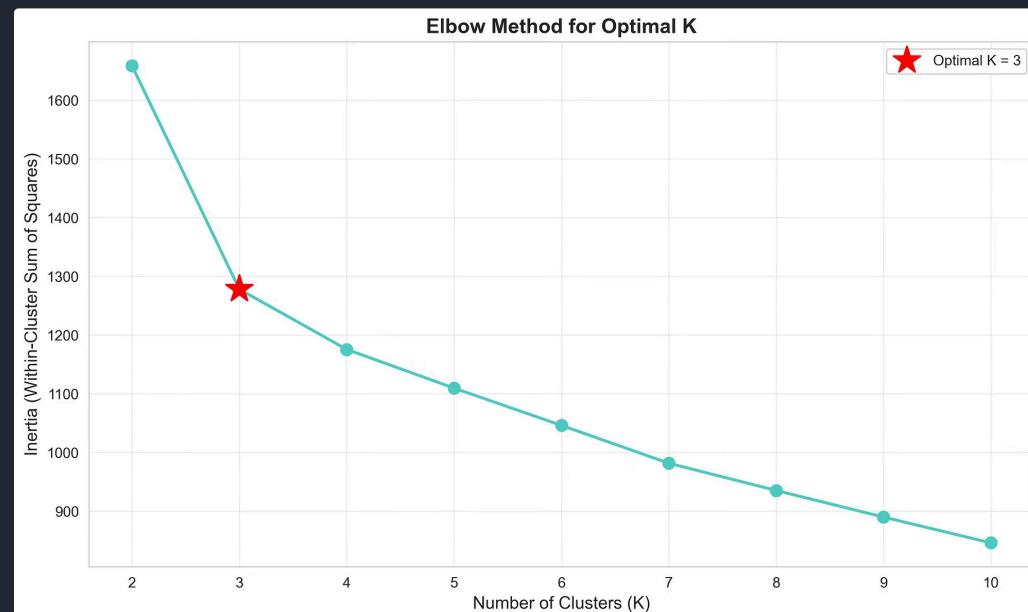
Évaluation

# Analyse K-Means

## Étape 1 - Déterminer le Nombre Optimal de Clusters

### Méthode du Coude

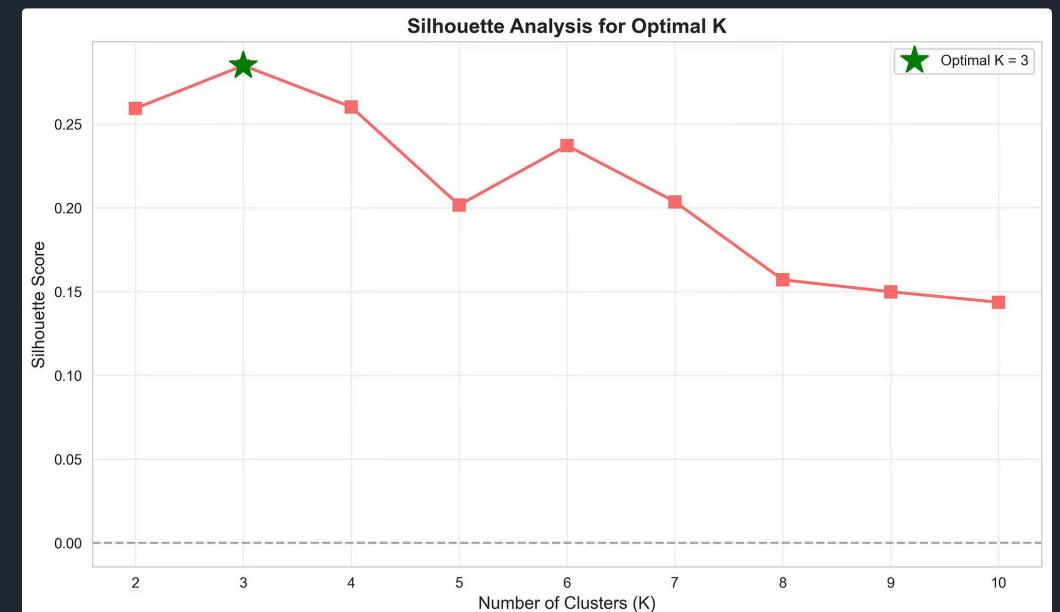
- "Trace l'inertie (compacité) vs nombre de clusters"
- "Le 'coude' où l'amélioration plafonne indique le K optimal"
- "Coude clair à K=3, correspondant au vrai nombre de cépages"



Légende : "Le coude à K=3 indique le nombre optimal de clusters"

### Analyse Silhouette

- "Mesure à quel point chaque point s'intègre dans son cluster"
- "Varie de -1 à 1, plus élevé = meilleur"
- "Pic à K=3 confirme la méthode du coude"



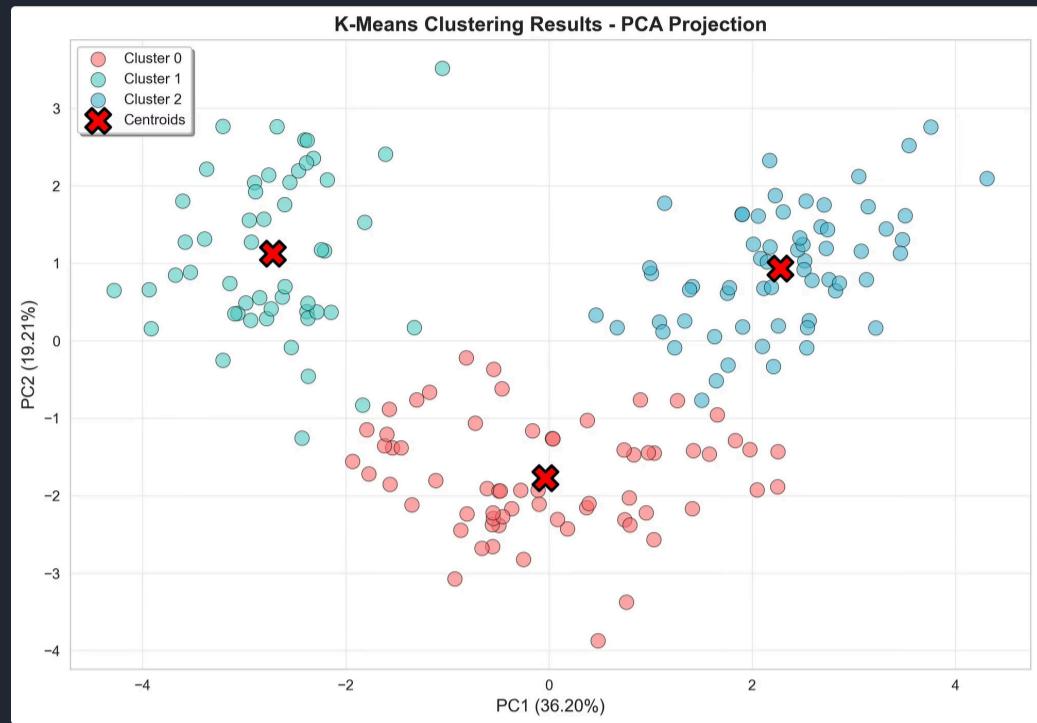
Légende : "Le score silhouette culmine à K=3, confirmant la méthode du coude"

# Analyse K-Means

## Résultats K-Means - Clusters de Haute Pureté

### Confirmation Visuelle

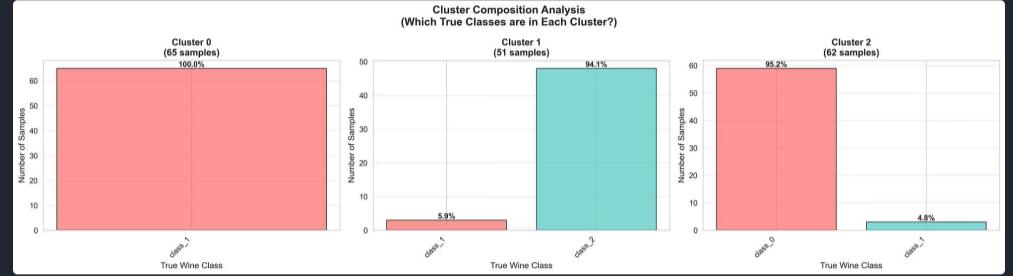
- La projection PCA montre une séparation spatiale claire
- Les croix rouges sont les centroïdes - le vin 'moyen' de chaque type"
- K-Means partitionne l'espace autour de ces centres



Légende : Clusters K-Means avec centroïdes (croix rouges)

### Qualité des Clusters

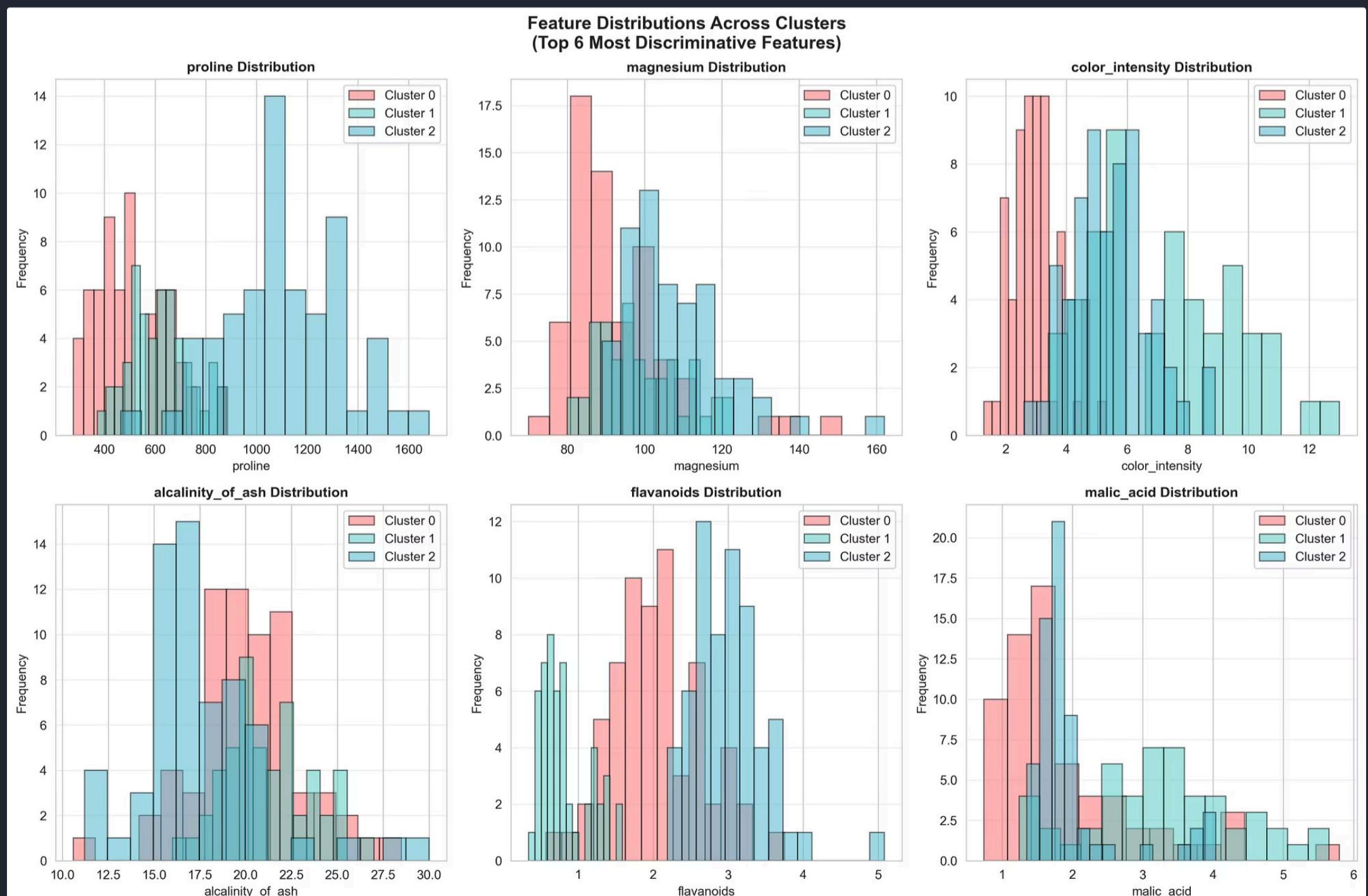
- K-Means a obtenu d'excellents résultats : 94-100% de pureté"
- Cluster 0 : 100% pur - a parfaitement capturé un cépage"
- Clusters 1 & 2 : 94-95% purs - mélange minimal"



Légende : Pureté des clusters : 94-100% d'affectations correctes

### Base Chimique

- Les distributions de features révèlent pourquoi les clusters diffèrent
- Exemple : Les niveaux de proline distinguent clairement le Cluster 2"
- Le contenu en magnésium sépare le Cluster 0
- Ce ne sont pas des groupes arbitraires - ils reflètent une vraie chimie



Légende : Les caractéristiques chimiques distinguent les clusters

100%

Cluster 0 Pureté

Parfaitement capturé un cépage

94%

Cluster 1 Pureté

Mélange minimal

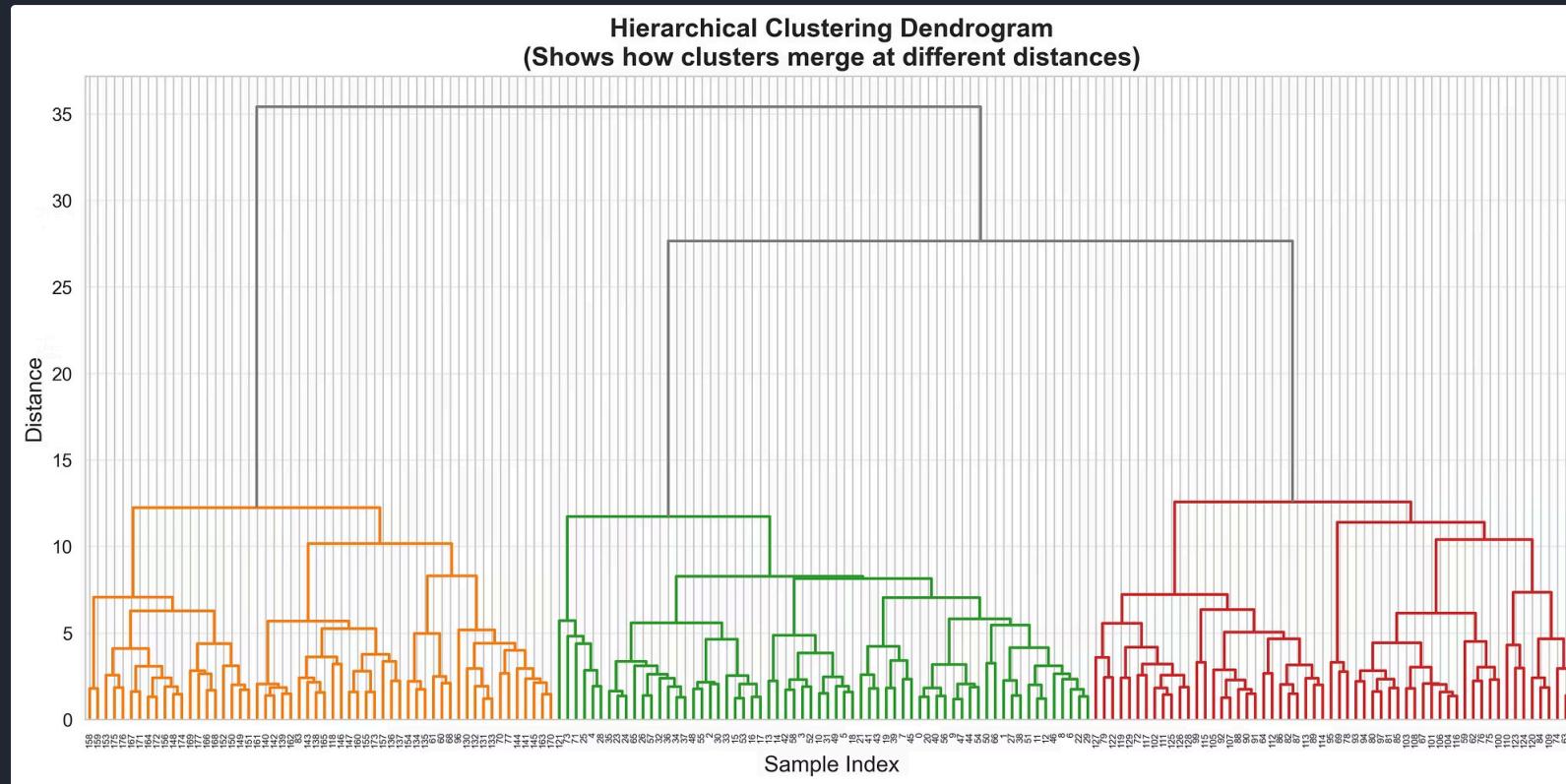
95%

Cluster 2 Pureté

Mélange minimal

# Analyse CAH

## Structure Hiérarchique et Performance



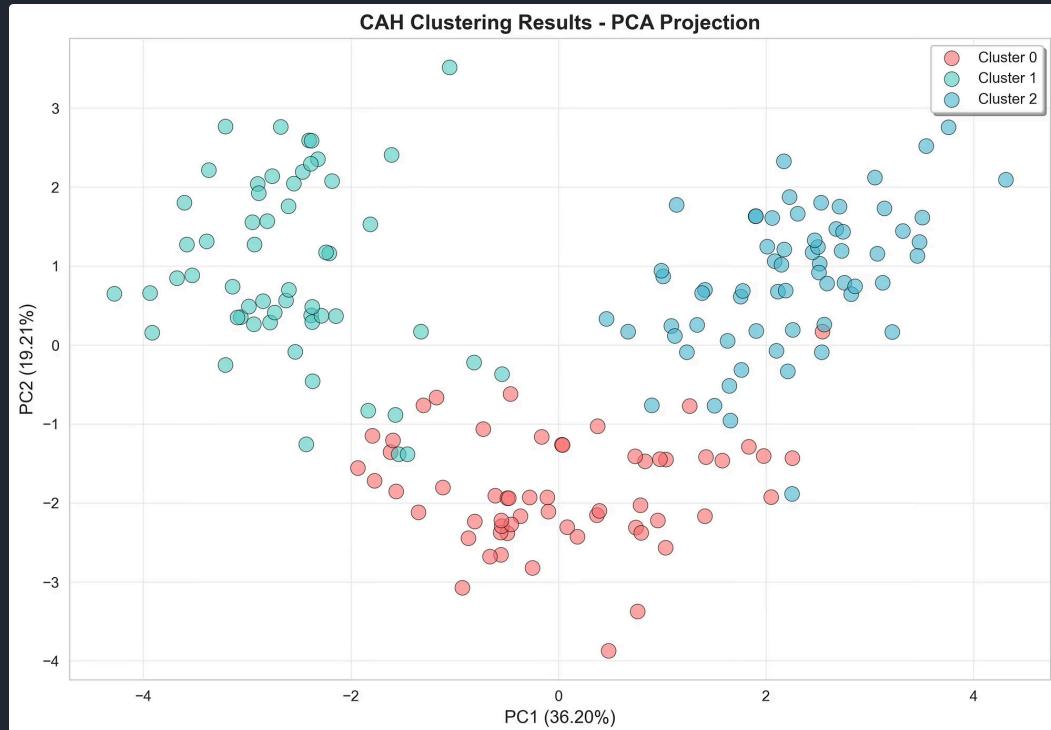
Légende : "Le dendrogramme révèle trois branches distinctes à distance ~27"

## Lecture du Dendrogramme

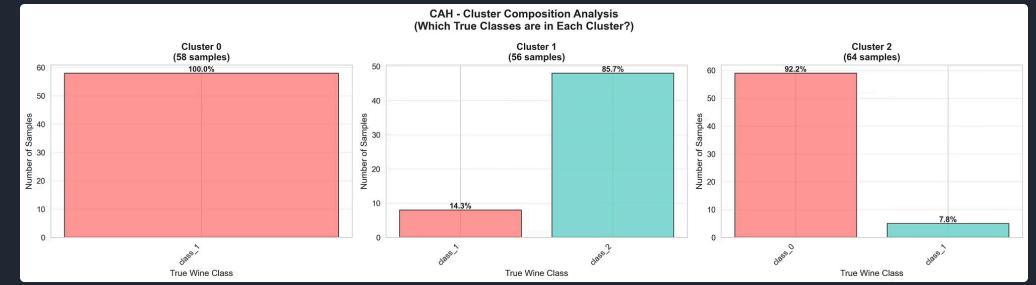
- Bas : vins individuels. Haut : tous fusionnés ensemble
- La hauteur indique la dissimilarité - fusion plus haute = plus différent
- Trois branches colorées émergent autour de la distance 27-28
- Cela confirme 3 comme nombre naturel de groupes

# CAH - Performance du Clustering

CAH Clustering Results - PCA Projection



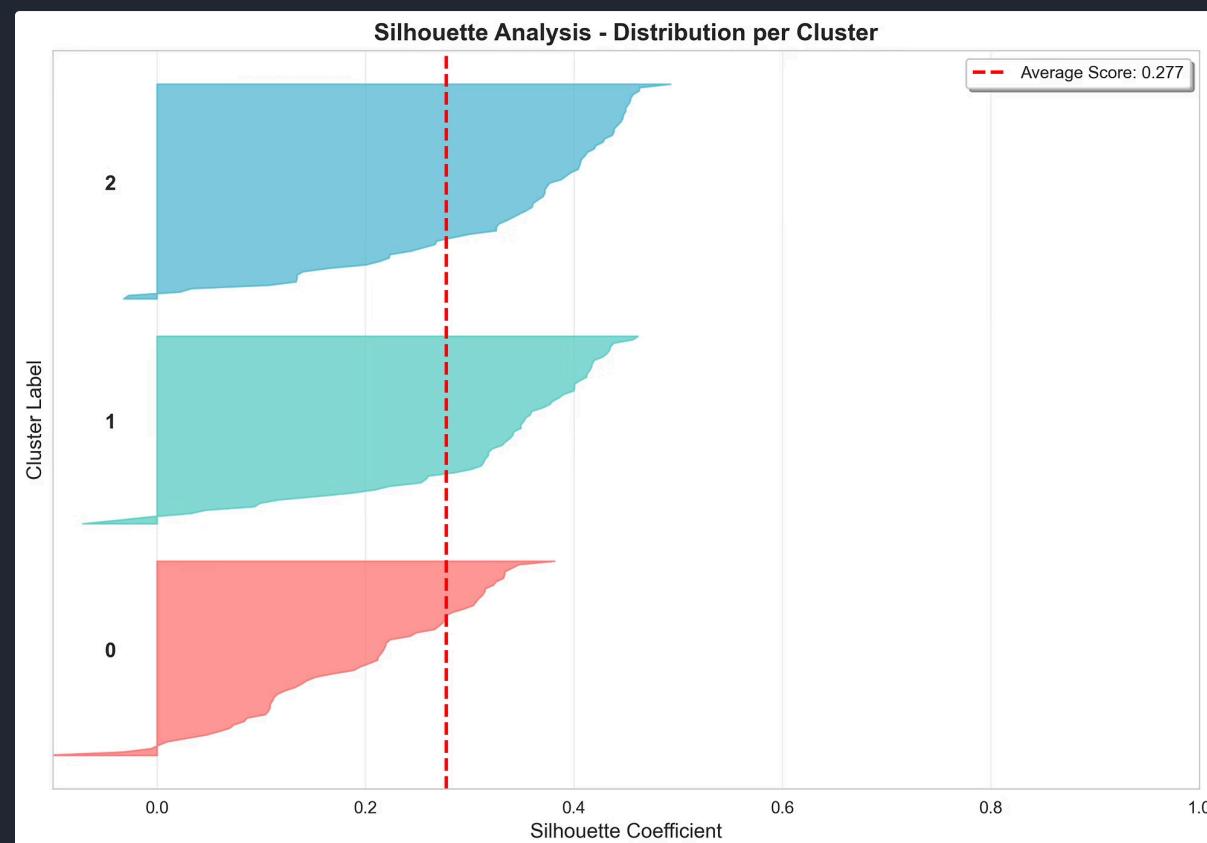
CAH - Cluster Composition Analysis



Légende : "Pureté des clusters : Cluster 0 = 100%, Cluster 1 = 86%, Cluster 2 = 92%"

Légende : "Résultats du clustering hiérarchique (CAH)"

Silhouette Analysis - Distribution per Cluster (CAH)

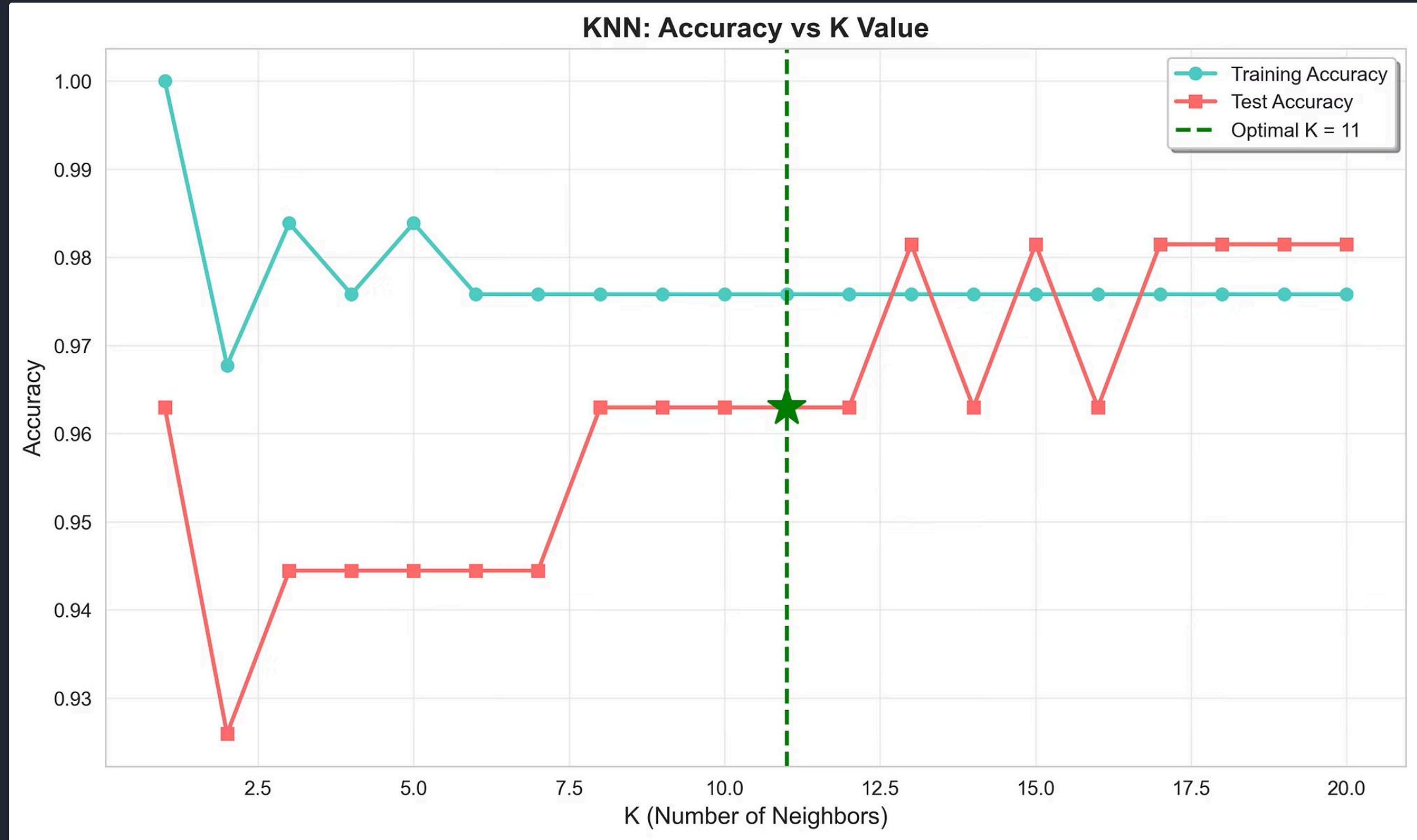


Légende : "Score silhouette moyen : 0,277"

Point à retenir : CAH offre l'avantage supplémentaire de la structure hiérarchique tout en atteignant des performances comparables à K-Means.

# Analyse KNN

## Optimiser K et Résultats de Classification



Légende : "K optimal=11 équilibre performances entraînement et test"

## Processus d'Optimisation

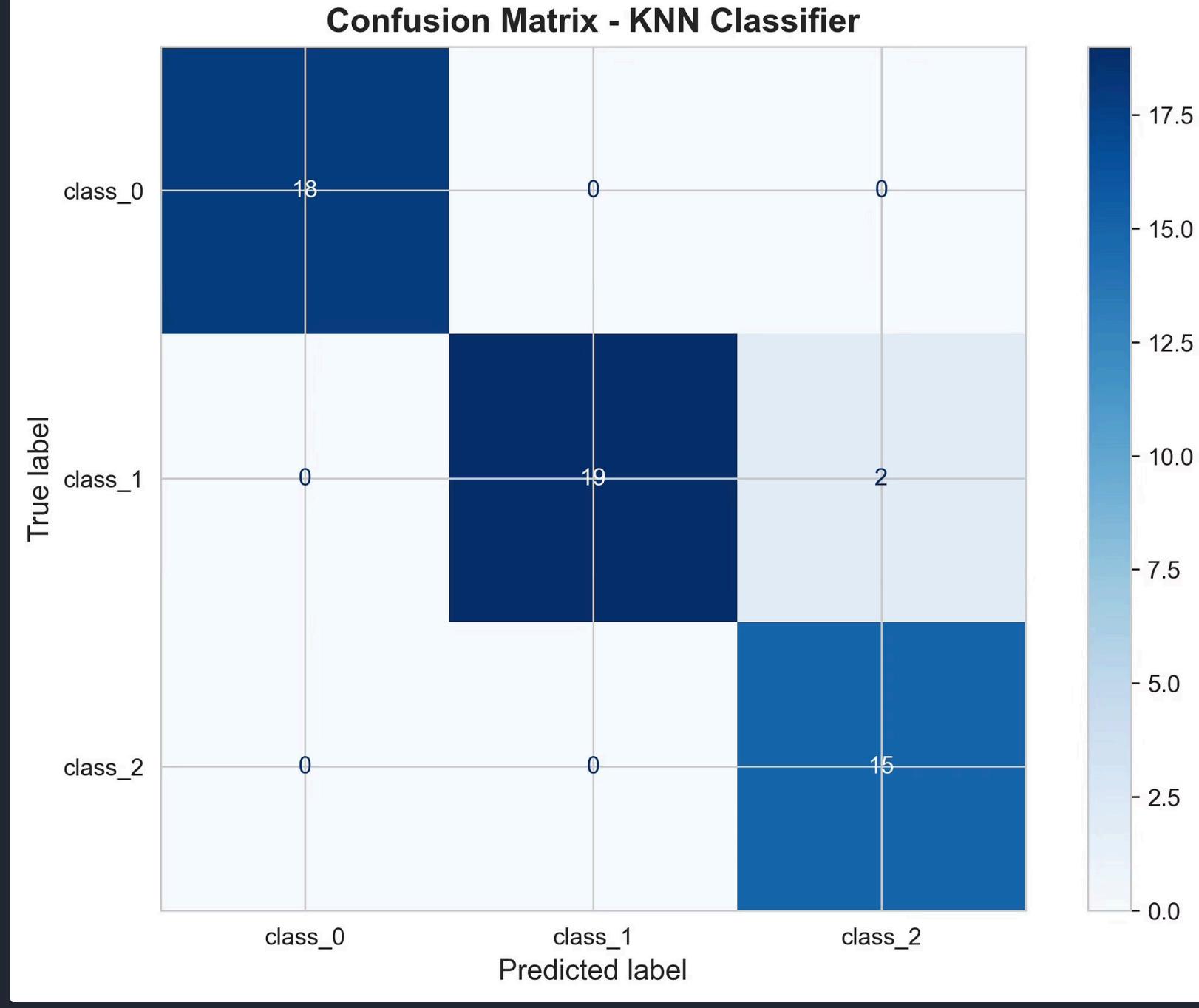
- "Testé K de 1 à 20 en utilisant la validation croisée"
- "K=1 : Surapprentissage - 100% entraînement mais précision test plus faible"
- "K=11 : Point optimal - équilibre précision et généralisation"
- "K>15 : Plateau - ajouter des voisins n'aide pas"

# Résultats de Classification

## Résultats KNN

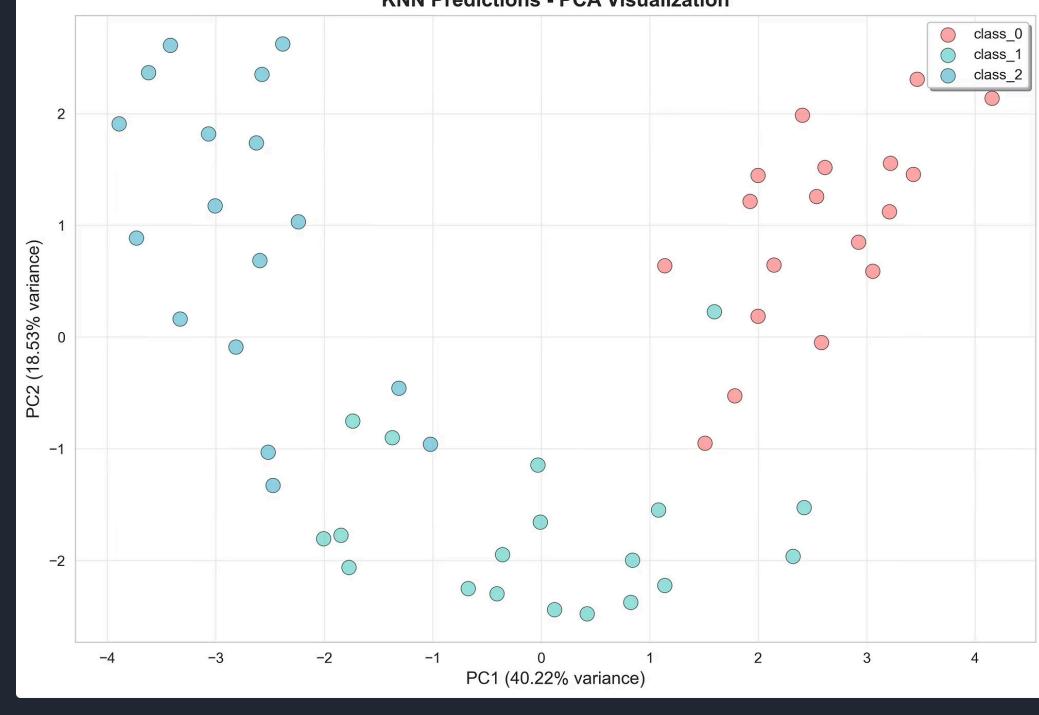
- Excellente Performance Prédictive
- 96,3% de précision test (52/54 correct)
- Précision parfaite sur les classes 0 et 2
- Haute confiance sur les prédictions correctes

## Matrice de Confusion - Classificateur KNN

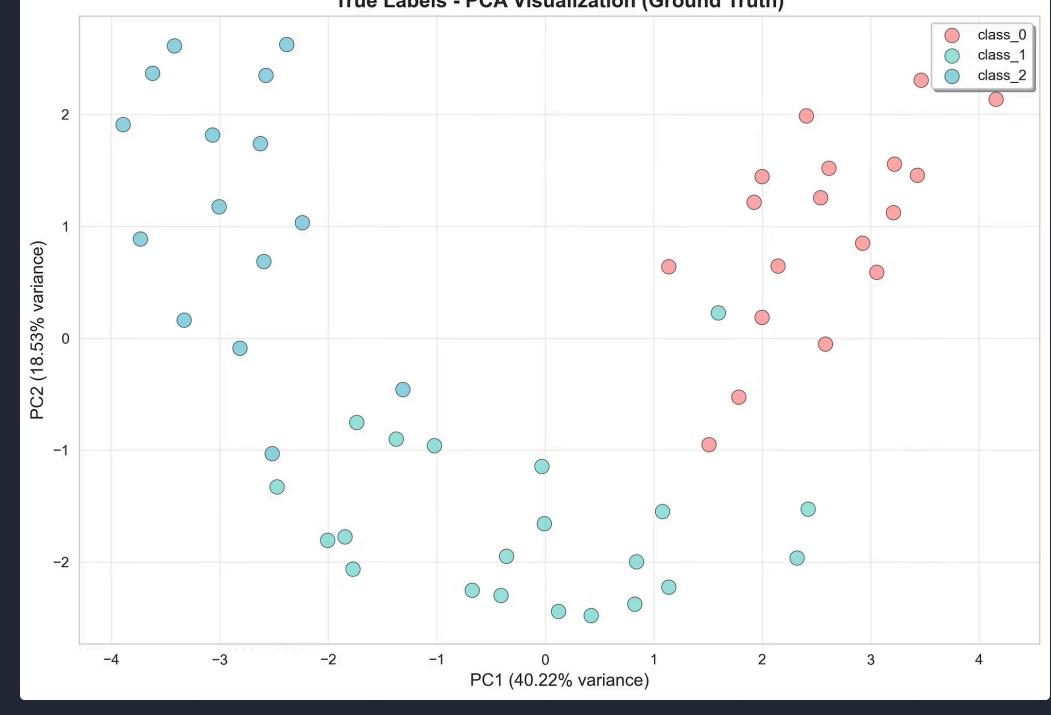


Légende : "96,3% de précision : seulement 2 vins classe\_1 mal classés"

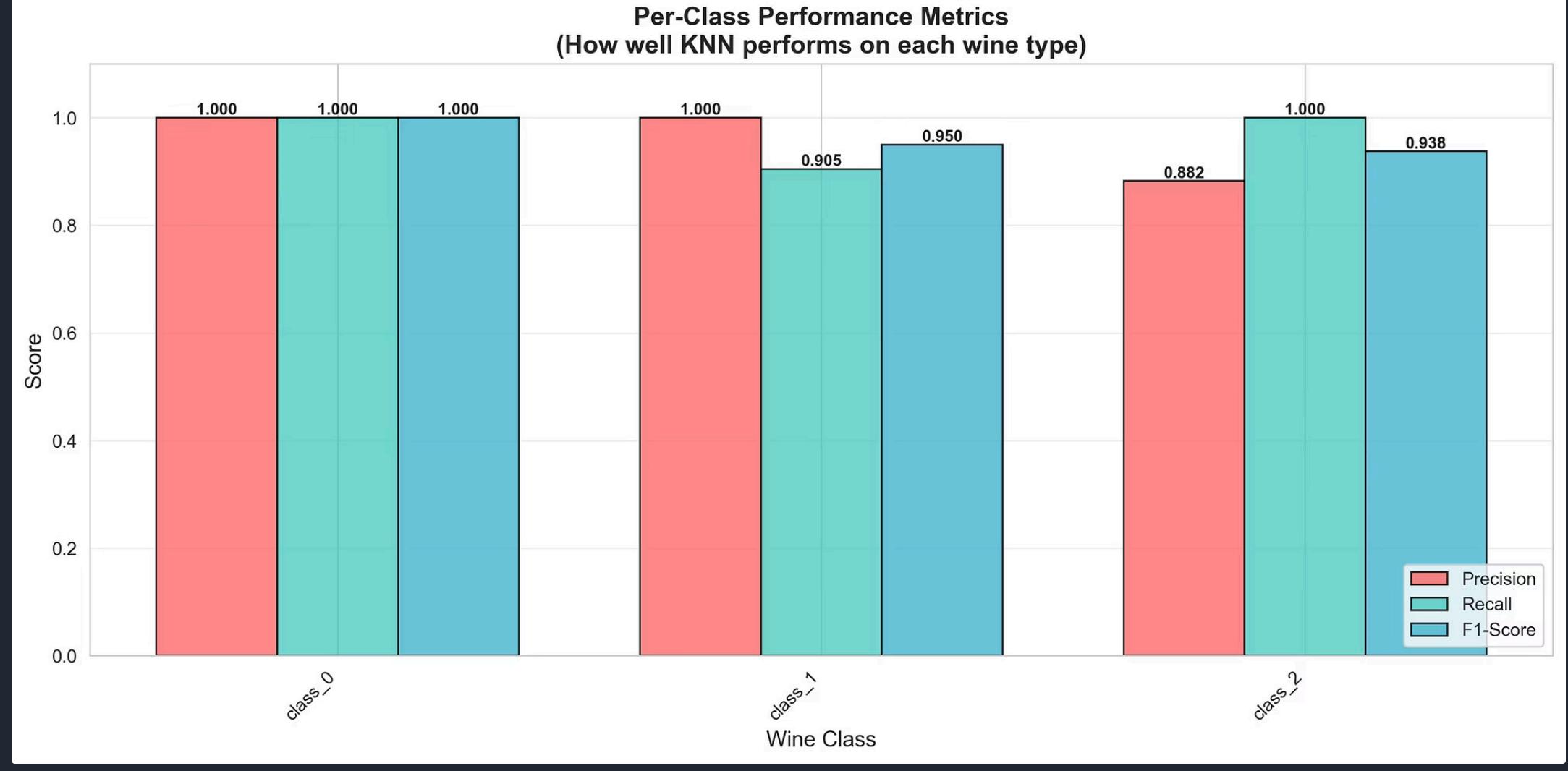
## Prédictions KNN



## Étiquettes Réelles



Légende : "Les étiquettes prédites vs réelles montrent un excellent accord"



Légende : "Performance équilibrée sur tous les types de vins"

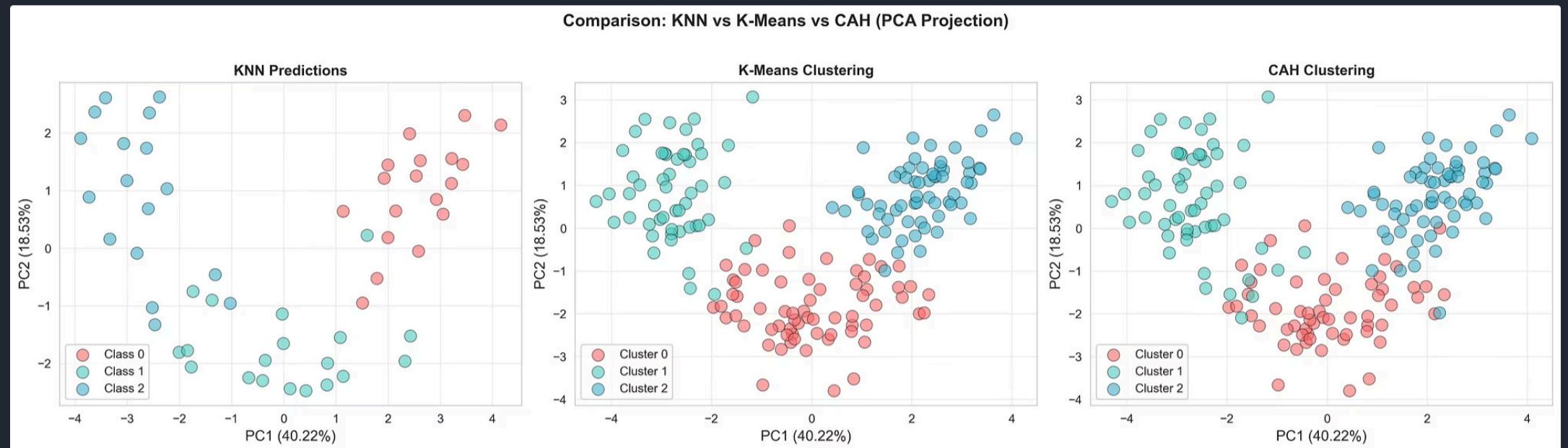
Conclusion forte : La précision de 96,3% de KNN et sa haute confiance sur les prédictions correctes le rendent très fiable pour la classification de vins en pratique.

# Analyse Comparative

## Comparaison Visuelle

Les graphiques PCA côté à côté montrent que toutes les méthodes ont découvert une structure similaire"

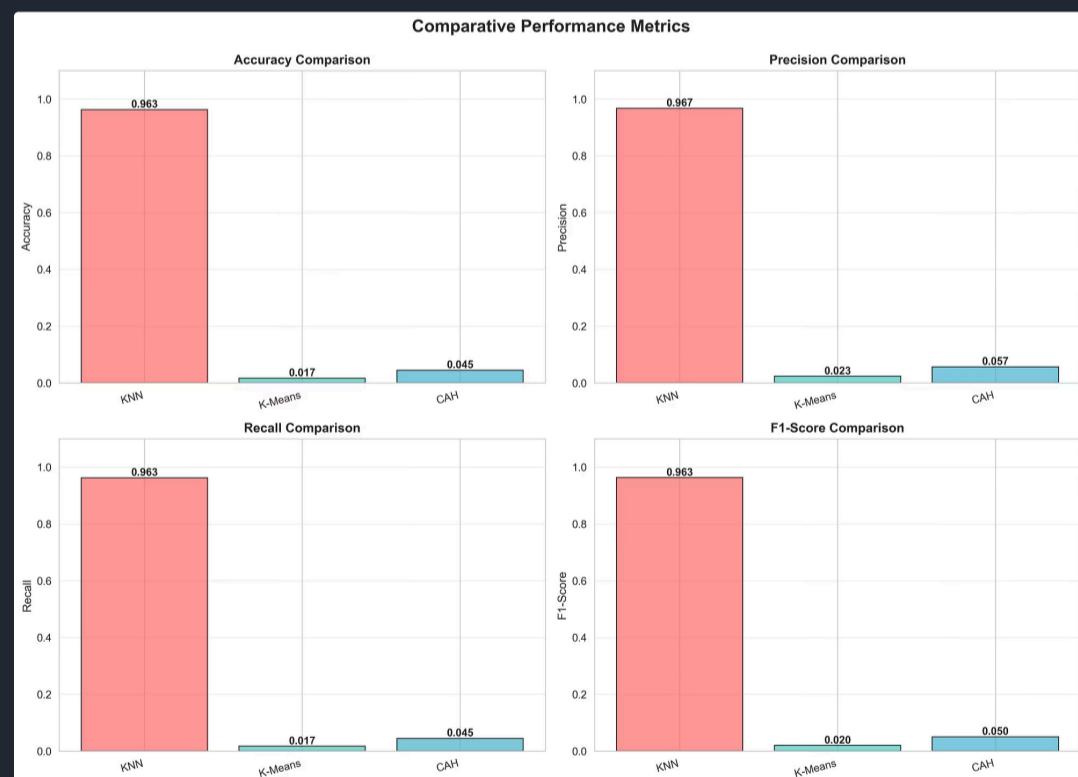
- "KNN : Frontières de décision claires"
- "K-Means : Clusters serrés avec centroïdes"
- "CAH : Groupements similaires avec contexte hiérarchique"



Légende : "Les trois méthodes ont identifié des groupements de vins similaires"

## Métriques de Performance Comparatives

K-Means et CAH montrent de faibles scores de 'précision' (~0,02-0,05) dans le graphique des métriques. C'est parce que nous les mesurons comme classificateurs (ce qu'ils ne sont pas). Leur vraie performance est montrée dans la pureté des clusters (94-100%).  
Métriques différentes pour tâches différentes !



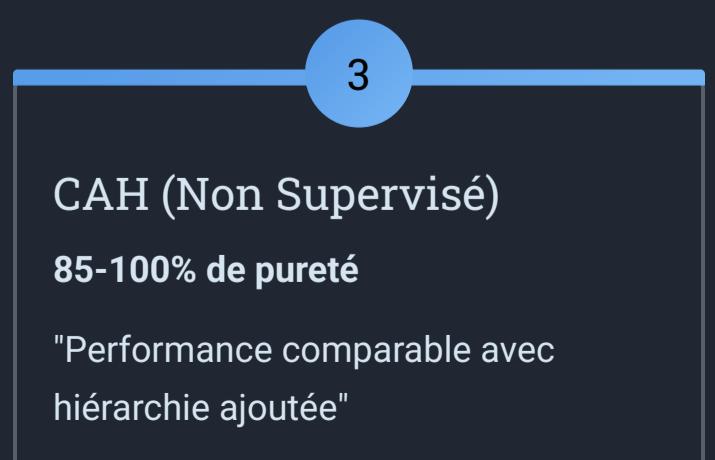
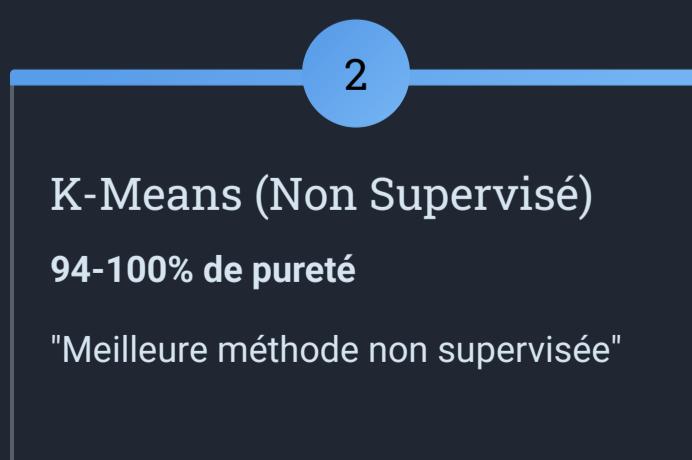
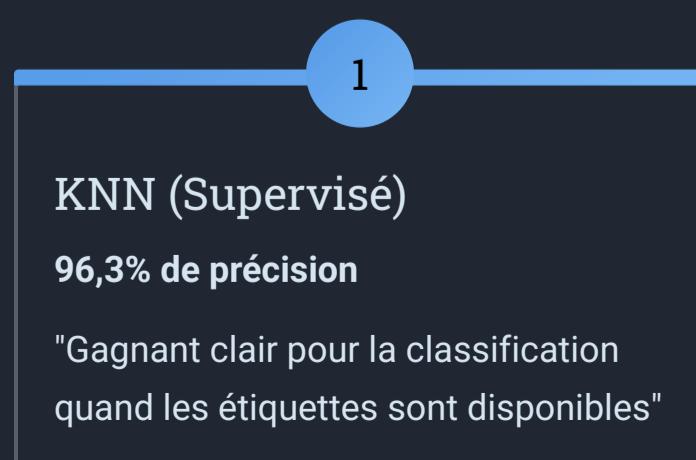
Légende : "KNN domine les métriques de classification ; les méthodes de clustering montrent des scores plus faibles (mesurant des tâches différentes)"

## Temps d'Exécution



Légende : "Toutes les méthodes complètes en moins de 10 secondes"

## Comparaison Directe des Performances



# Conclusions & Perspectives

## Principales Découvertes

### 1 Les trois méthodes ont analysé avec succès le dataset de vins

- "KNN : 96,3% de précision de classification"
- "K-Means : 94-100% de pureté de clusters"
- "CAH : 85-100% de pureté de clusters avec hiérarchie"

### 2 Structure du dataset confirmée

- "Les cépages ont des signatures chimiques distinctes"
- "Proline, magnésium et flavonoïdes sont des discriminants clés"
- "La séparation naturelle rend efficaces les méthodes supervisées et non supervisées"

### 3 La validation croisée renforce les résultats

- "Plusieurs méthodes d'optimisation concordent (coude + silhouette → K=3)"
- "K-Means et CAH ont identifié des patterns chimiques similaires"
- "La cohérence entre algorithmes valide les résultats"

## Quand Utiliser Chaque Méthode

### KNN

"Quand vous avez des données étiquetées et besoin de prédictions"

- Authentification de vins
- Contrôle qualité en production

### K-Means

"Quand vous avez besoin de clustering rapide et scalable"

- Segmentation clients
- Grands datasets

### CAH

"Quand les relations importent"

- Comprendre les similarités de vins
- Systèmes de taxonomie/classification

# Questions & Discussion

Merci

## Applications Réelles

- "Authentification de vins : Déetecter les contrefaçons"
- "Optimisation de vignobles : Associer chimie du sol aux cépages"
- "Prédiction de qualité : Prévoir caractéristiques du vin depuis chimie du raisin"
- "Segmentation marché : Grouper vins pour marketing"

## Leçons Apprises

- "La standardisation des features est critique pour les méthodes basées sur la distance"
- "Plusieurs méthodes de validation préviennent le surapprentissage"
- "Les méthodes non supervisées peuvent redécouvrir les classifications supervisées"
- "La visualisation (PCA) est essentielle pour comprendre les données haute dimension"