

PROJET MACHINE LEARNING

**Analyse de Clustering des Profils
Chimiques des Vins :
Identification de Groupes Basés sur la
Composition**

RÉALISÉ PAR:

Marwane KASSA
Fadoua BEN KABOUR
Kaoutar ABIDI
Ikrame HAFSI
Ilyas OUHNINE

ENCADRÉ PAR:

Youssef LAMRANI

PLAN

1

2

3

4

5

6

Définition du problème

Présentation des algorithmes

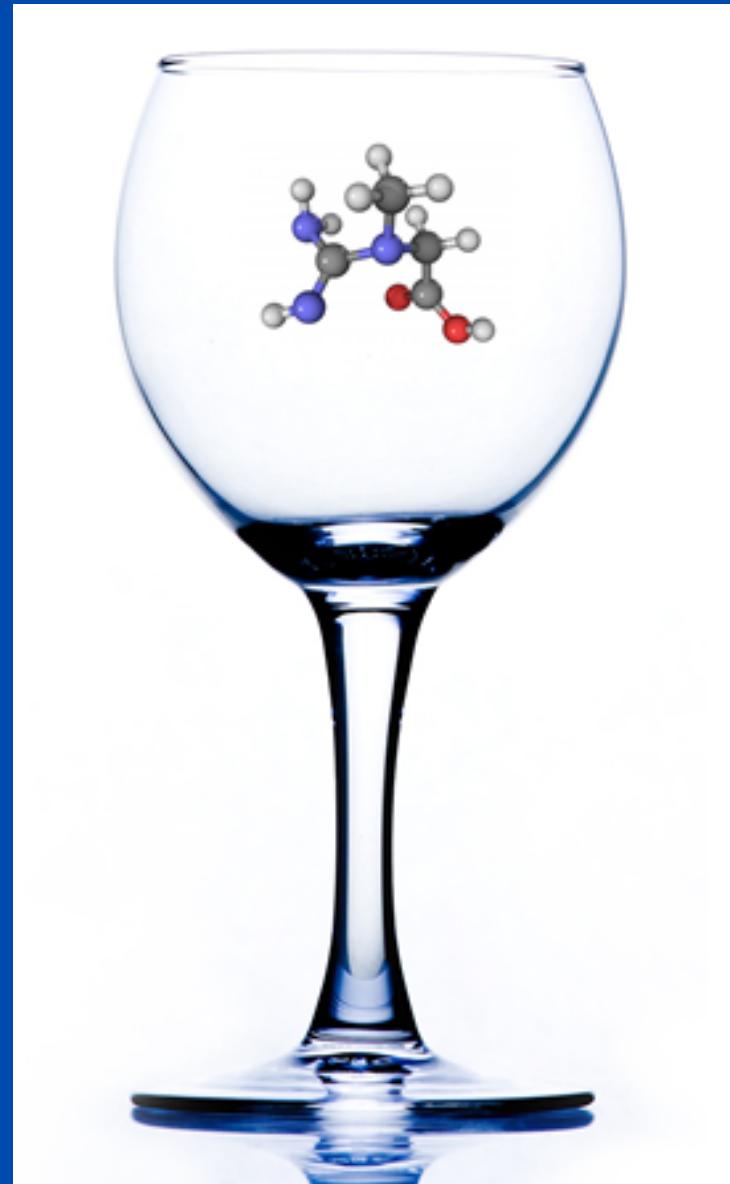
**Modélisation et implémentation
du modèle**

Analyse des Résultats

Application déployée

Limites et améliorations

PROBLÈME & OBJECTIFS



Formulation du Problème :

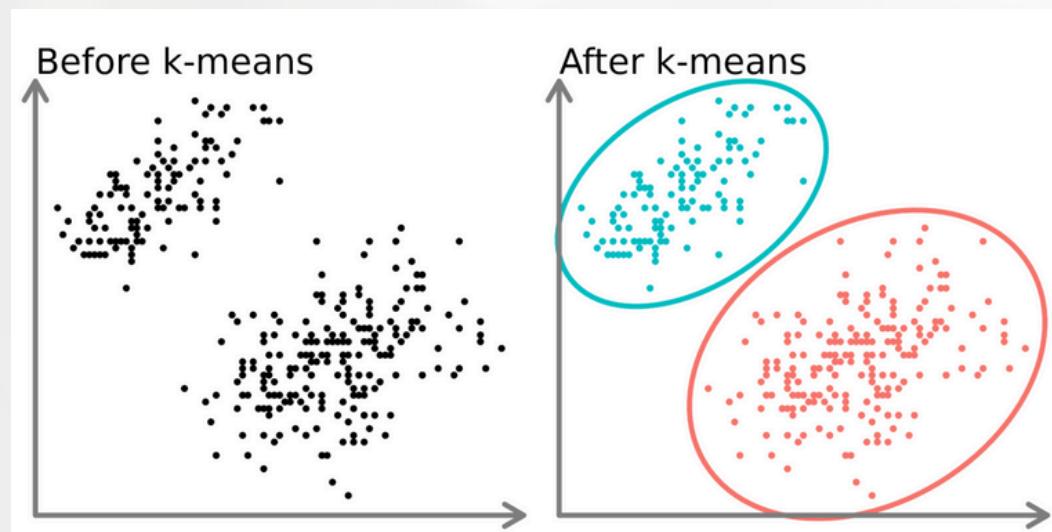
Comment regrouper les vins en fonction de leurs caractéristiques chimiques sans utiliser d'étiquettes prédéfinies, afin de révéler des clusters naturels et de détecter des similitudes ou divergences inattendues ?

Objectifs :

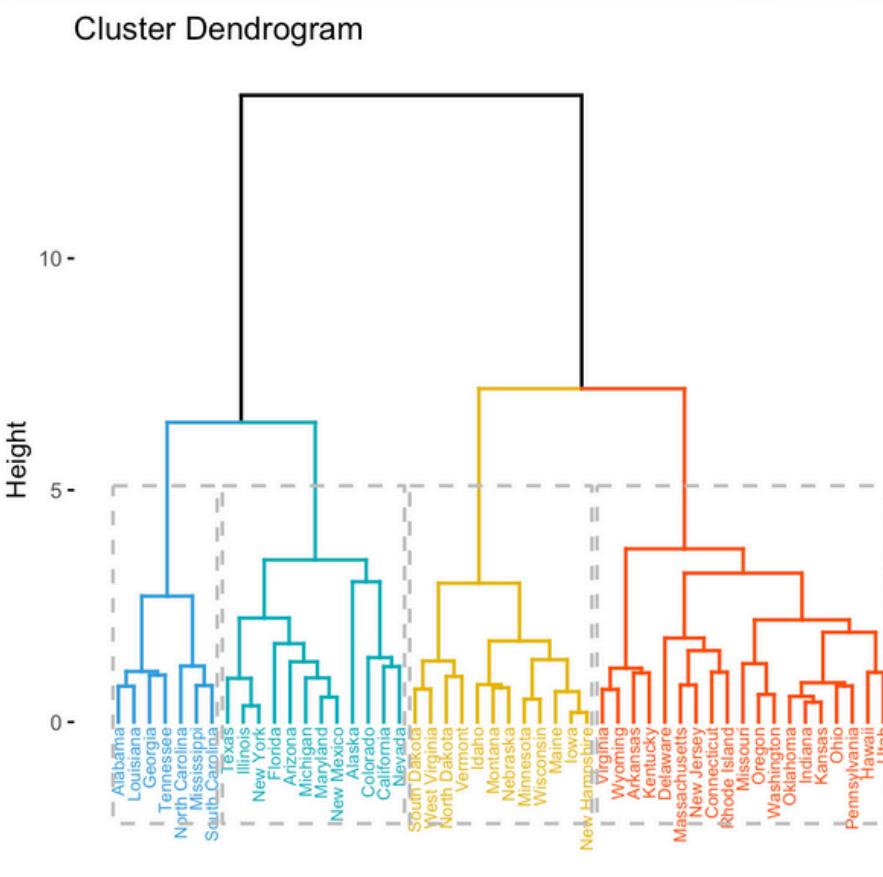
- Utiliser le clustering pour identifier des regroupements naturels
- Découvrir des similitudes ou divergences sans étiquettes préexistantes
- Explorer si les catégories traditionnelles se manifestent naturellement

PRESENTATION DES ALGORITHMES

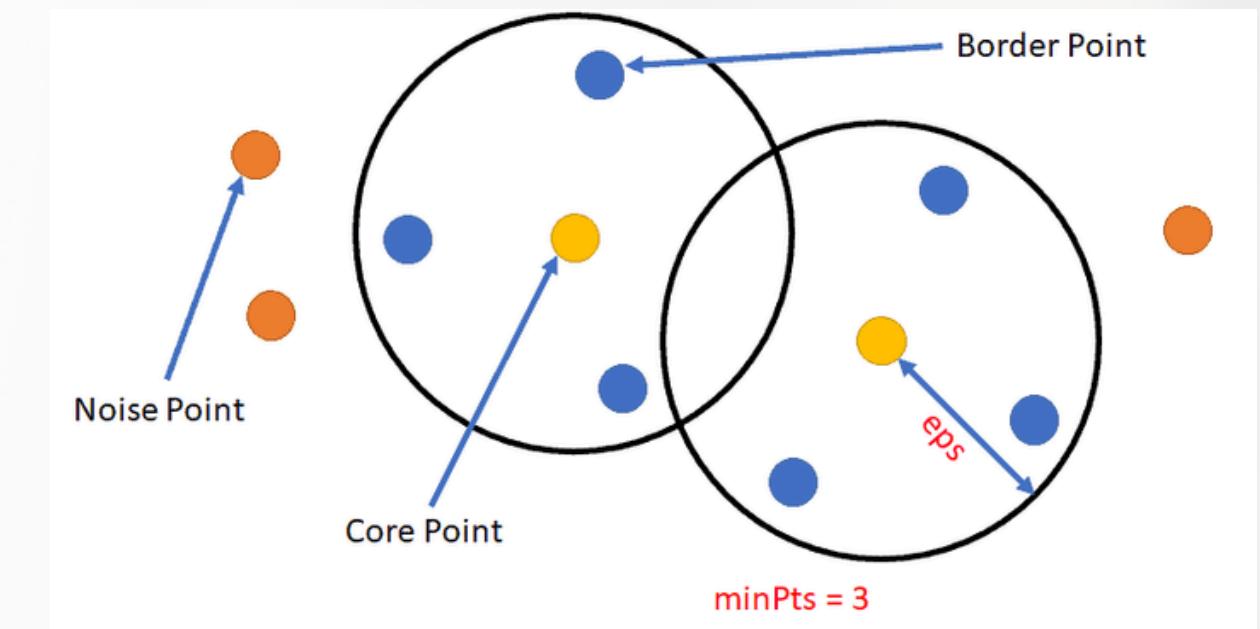
K-means



CAH



DBSCAN



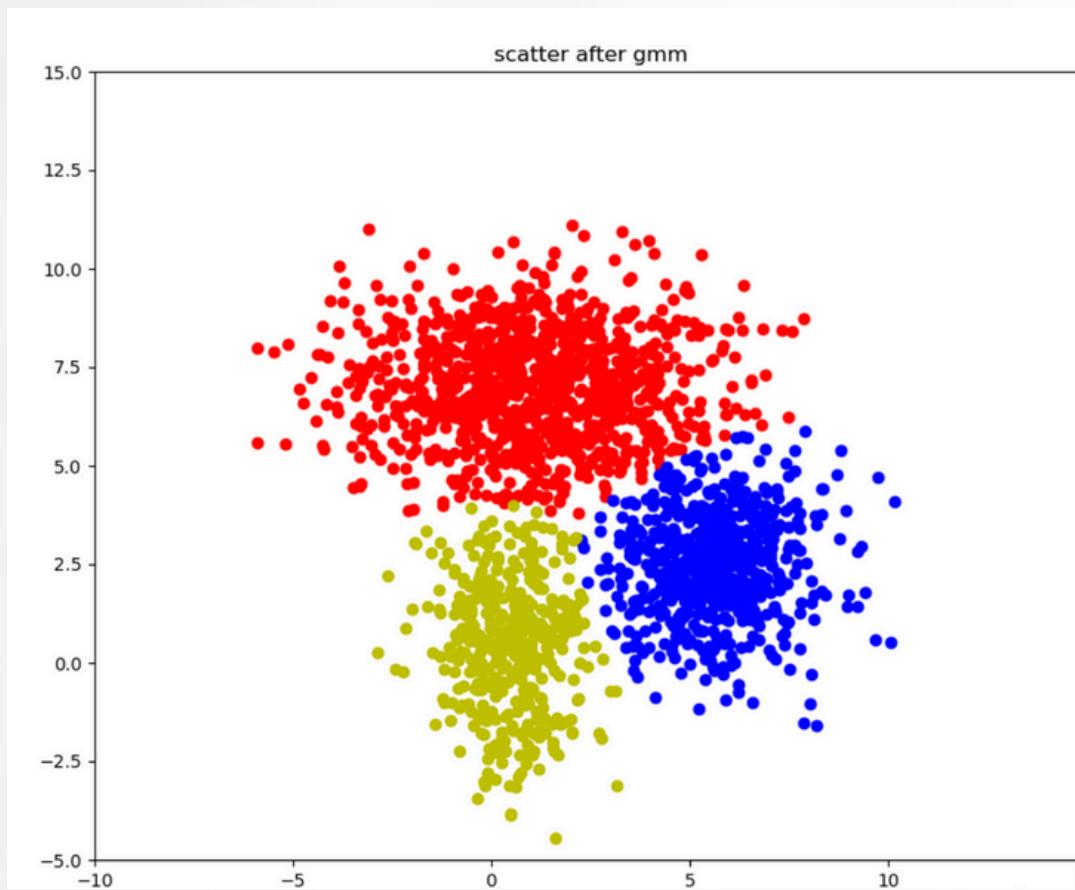
- + Simple et rapide à implémenter
- Sensible au choix des points initiaux

- + Pas besoin de fixer le nombre de clusters à l'avance
- Complexité élevée pour de grands jeux de données

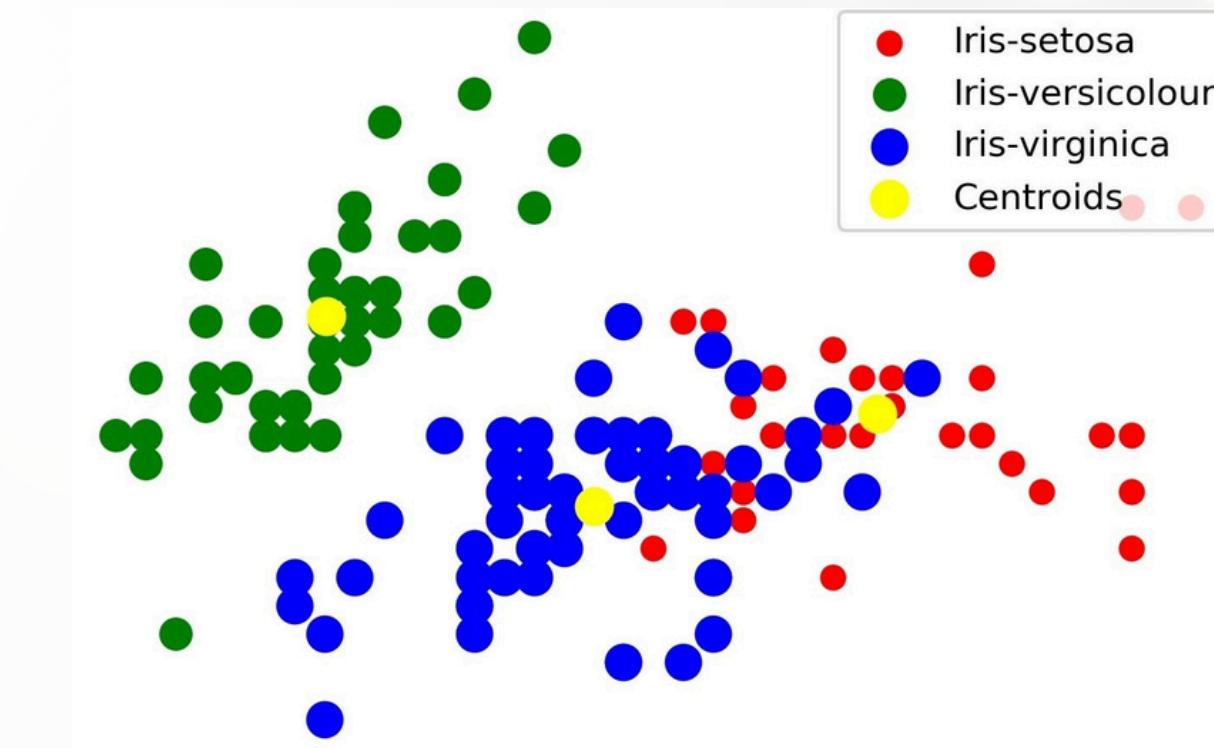
- + Capable de détecter des clusters de formes variées
- Sensible aux paramètres (*et MinPts*)

PRESENTATION DES ALGORITHMES

GMM



BIRCH



+Simple et rapide à implémenter
-Nécessite de spécifier le
nombre de clusters

+ Adapté pour des bases de
données volumineuses
- Limité aux clusters de formes
ellipsoïdales

EXPLORATION ET PRÉPARATION DES DONNÉES

Exploration et Préparation des Données

```
▶ print(df.head()) # View the first 5 rows
```

```
→   Class Alcohol Malic_Acid Ash Ash_Alcanity Magnesium Total_Phenols \
0    1     14.23      1.71  2.43        15.6       127        2.80
1    1     13.20      1.78  2.14        11.2       100        2.65
2    1     13.16      2.36  2.67        18.6       101        2.80
3    1     14.37      1.95  2.50        16.8       113        3.85
4    1     13.24      2.59  2.87        21.0       118        2.80

   Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity Hue \
0      3.06          0.28           2.29          5.64  1.04
1      2.76          0.26           1.28          4.38  1.05
2      3.24          0.30           2.81          5.68  1.03
3      3.49          0.24           2.18          7.80  0.86
4      2.69          0.39           1.82          4.32  1.04

   OD280  Proline
0    3.92     1065
1    3.40     1050
2    3.17     1185
3    3.45     1480
4    2.93      735
```

perçu des premières lignes du jeu de données des vins

```
▶ df = df.iloc[:, 1:]
```

```
print(df.head())
→   Alcohol Malic_Acid Ash Ash_Alcanity Magnesium Total_Phenols \
0     14.23      1.71  2.43        15.6       127        2.80
1     13.20      1.78  2.14        11.2       100        2.65
2     13.16      2.36  2.67        18.6       101        2.80
3     14.37      1.95  2.50        16.8       113        3.85
4     13.24      2.59  2.87        21.0       118        2.80

   Flavanoids Nonflavanoid_Phenols Proanthocyanins Color_Intensity Hue \
0      3.06          0.28           2.29          5.64  1.04
1      2.76          0.26           1.28          4.38  1.05
2      3.24          0.30           2.81          5.68  1.03
3      3.49          0.24           2.18          7.80  0.86
4      2.69          0.39           1.82          4.32  1.04

   OD280  Proline
0    3.92     1065
1    3.40     1050
2    3.17     1185
3    3.45     1480
4    2.93      735
```

préparation pour le clustering)

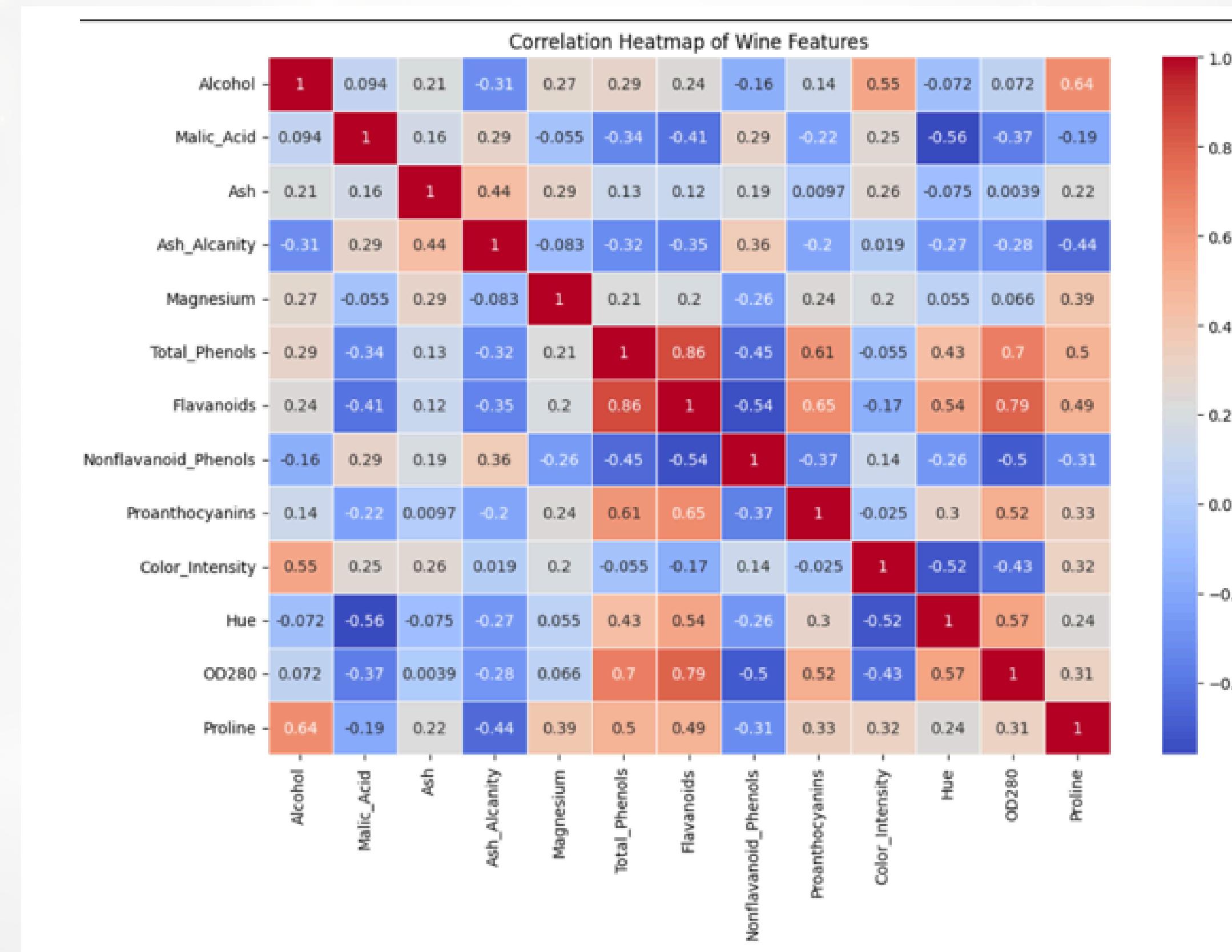
EXPLORATION ET PRÉPARATION DES DONNÉES

Statistiques Descriptives des Variables Chimiques du Vin

	count	mean	std	min	25%	50%	75%	max	skewness	Kurtosis
Alcohol	178.0	13.000618	0.811827	11.03	12.3625	13.050	13.6775	14.83	-0.051482	-0.852500
Malic_Acid	178.0	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.80	1.039651	0.299207
Ash	178.0	2.366517	0.274344	1.36	2.2100	2.360	2.5575	3.23	-0.176699	1.143978
Ash_Alcanity	178.0	19.494944	3.339564	10.60	17.2000	19.500	21.5000	30.00	0.213047	0.487942
Magnesium	178.0	99.741573	14.282484	70.00	88.0000	98.000	107.0000	162.00	1.098191	2.104991
Total_Phenols	178.0	2.295112	0.625851	0.96	1.7425	2.355	2.8000	3.88	0.086639	-0.835627
Flavanoids	178.0	2.029270	0.998859	0.34	1.2050	2.135	2.8750	5.08	0.025344	-0.880382
Nonflavanoid_Phenols	178.0	0.361854	0.124453	0.13	0.2700	0.340	0.4375	0.66	0.450151	-0.637191
Proanthocyanins	178.0	1.590899	0.572359	0.41	1.2500	1.555	1.9500	3.58	0.517137	0.554649
Color_Intensity	178.0	5.058090	2.318286	1.28	3.2200	4.690	6.2000	13.00	0.868585	0.381522
Hue	178.0	0.957449	0.228572	0.48	0.7825	0.965	1.1200	1.71	0.021091	-0.344096
OD280	178.0	2.611685	0.709990	1.27	1.9375	2.780	3.1700	4.00	-0.307285	-1.086435
Proline	178.0	746.893258	314.907474	278.00	500.5000	673.500	985.0000	1680.00	0.767822	-0.248403

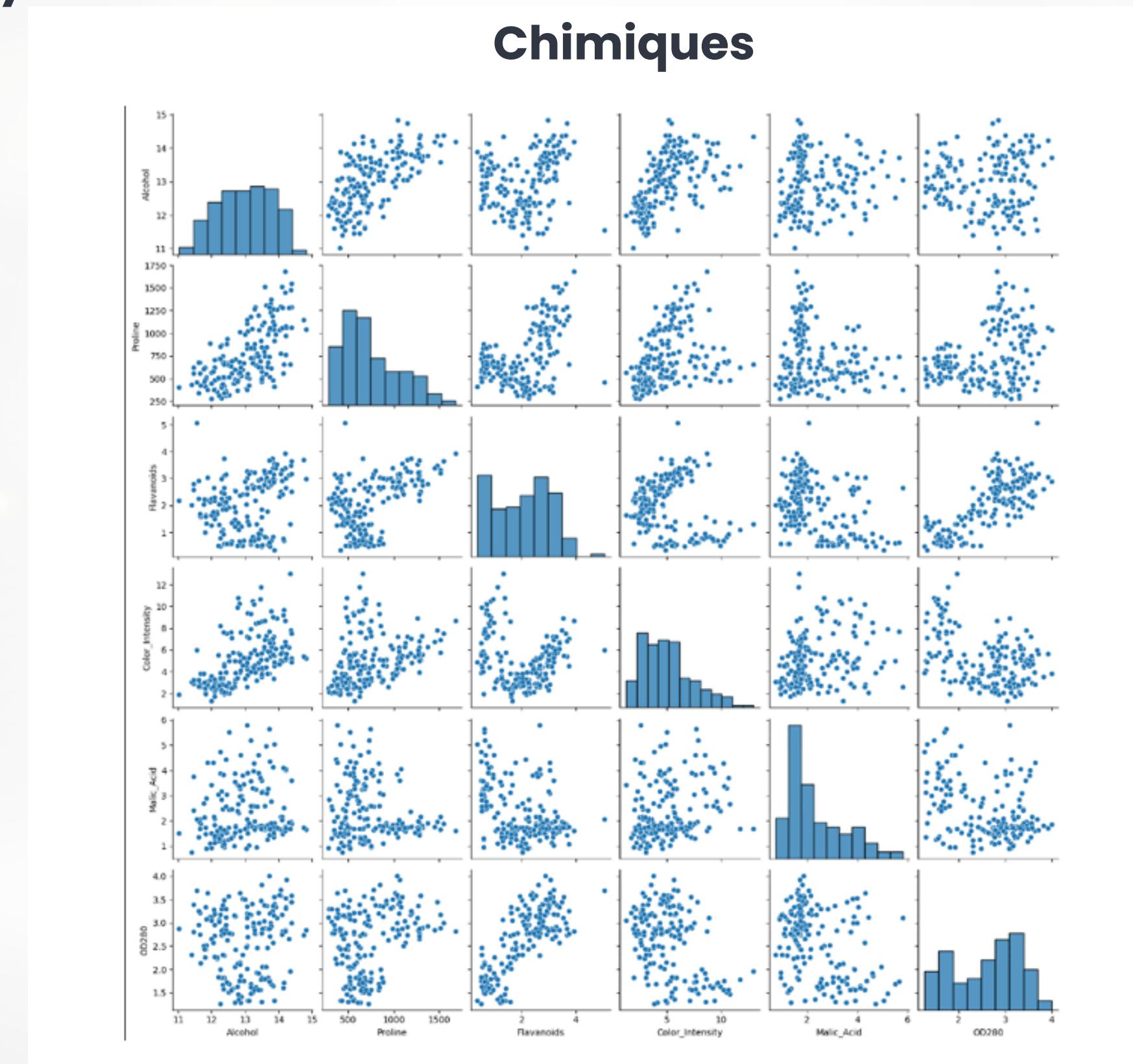
EXPLORATION ET PRÉPARATION DES DONNÉES

Matrice de Corrélation des Caractéristiques Chimiques du Vin



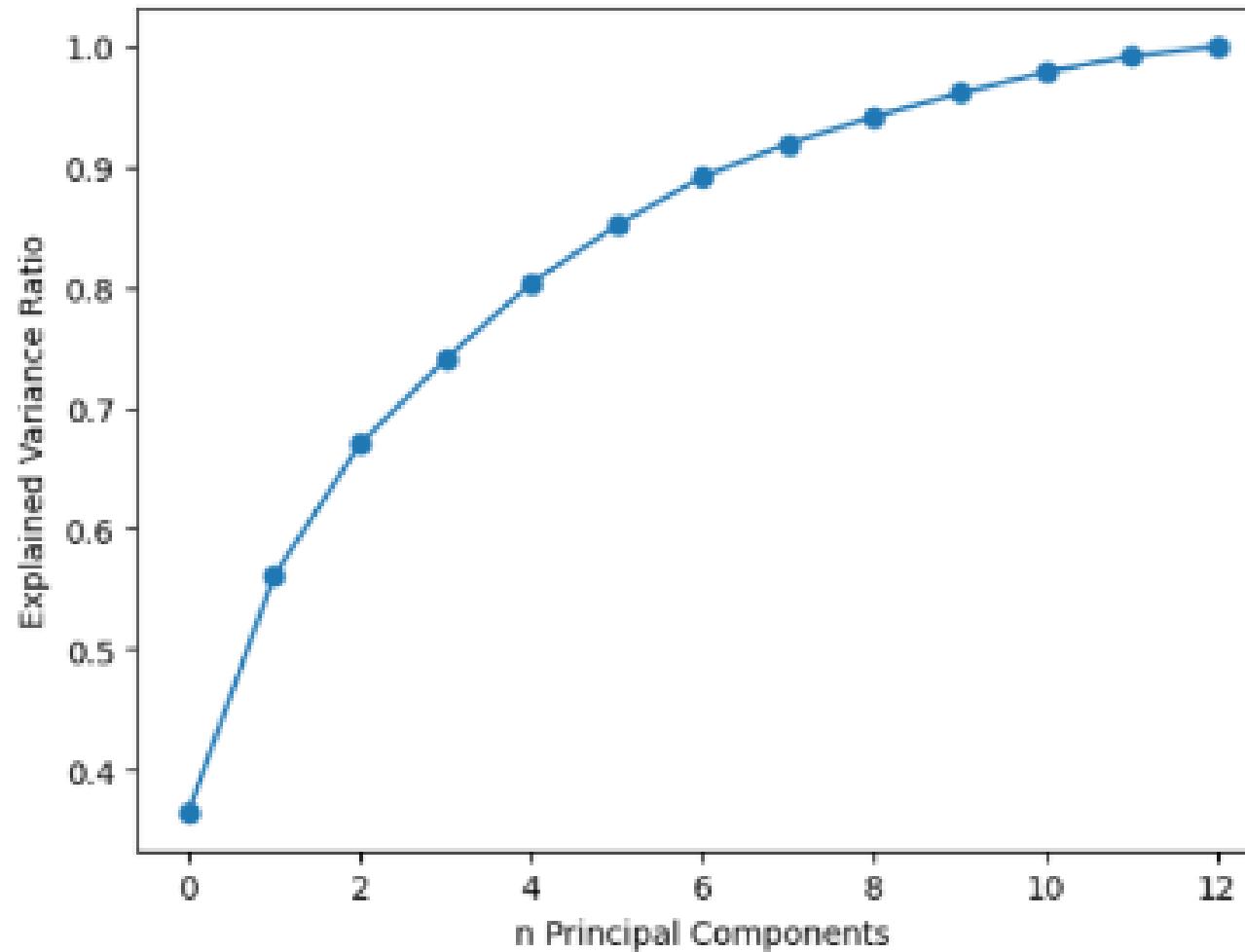
EXPLORATION ET PRÉPARATION DES DONNÉES

Analyse des Relations Bivariées et Distributions des Variables Chimiques



EXPLORATION ET PRÉPARATION DES DONNÉES

Diagramme en boîte de la concentration de flavonoïdes dans les vins



- Identification des composantes principales expliquant le plus de variance.
- Sélection des dimensions optimales pour réduire la complexité des données.
- Préservation de l'essentiel de l'information avec moins de variables.
- Facilitation du clustering en éliminant les dimensions non pertinentes.
- Assurance d'une meilleure interprétation des résultats en réduisant le bruit.

MODÉLISATION ET IMPLÉMENTATION DU MODÈLE

Pour chaque algorithme, une recherche de grille a été réalisée pour ajuster les principaux paramètres:

	Model	Best Parameters	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	KMeans	{'init': 'k-means++', 'max_iter': 300, 'n_clus...}	0.345719	1.151019	99.056198
1	DBSCAN	{'eps': 0.5, 'min_samples': 2}	0.788587	0.207630	191.836471
2	AgglomerativeClustering	{'linkage': 'ward', 'n_clusters': 3}	0.340378	1.171293	96.991639
3	GaussianMixture	{'covariance_type': 'tied', 'n_components': 3}	0.344133	1.162272	97.226104
4	BIRCH	{'n_clusters': 3, 'threshold': 0.5}	0.338897	1.157332	96.026209

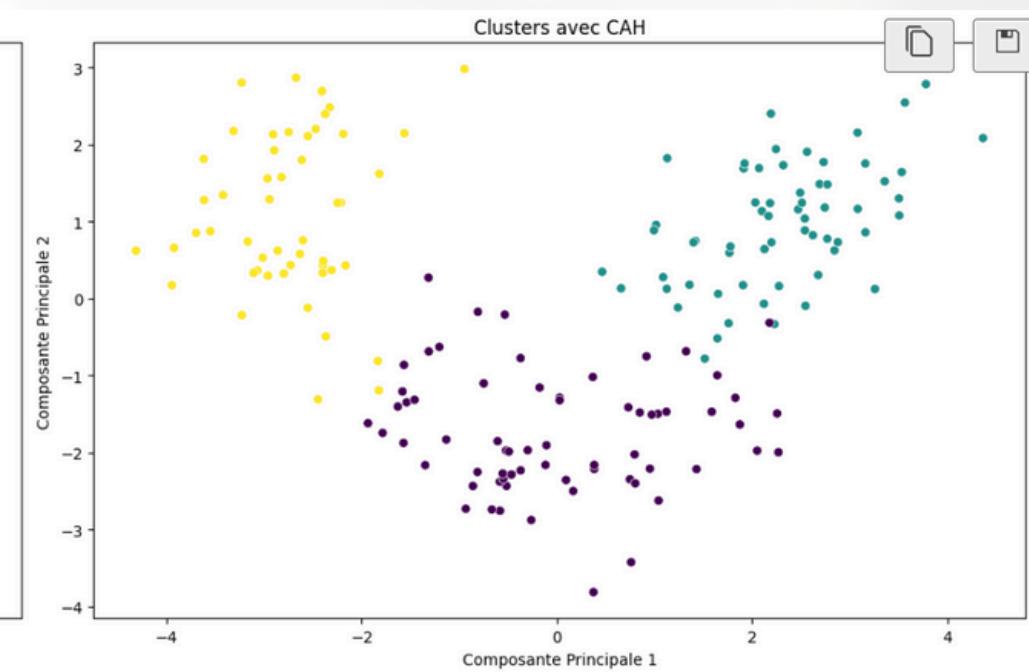
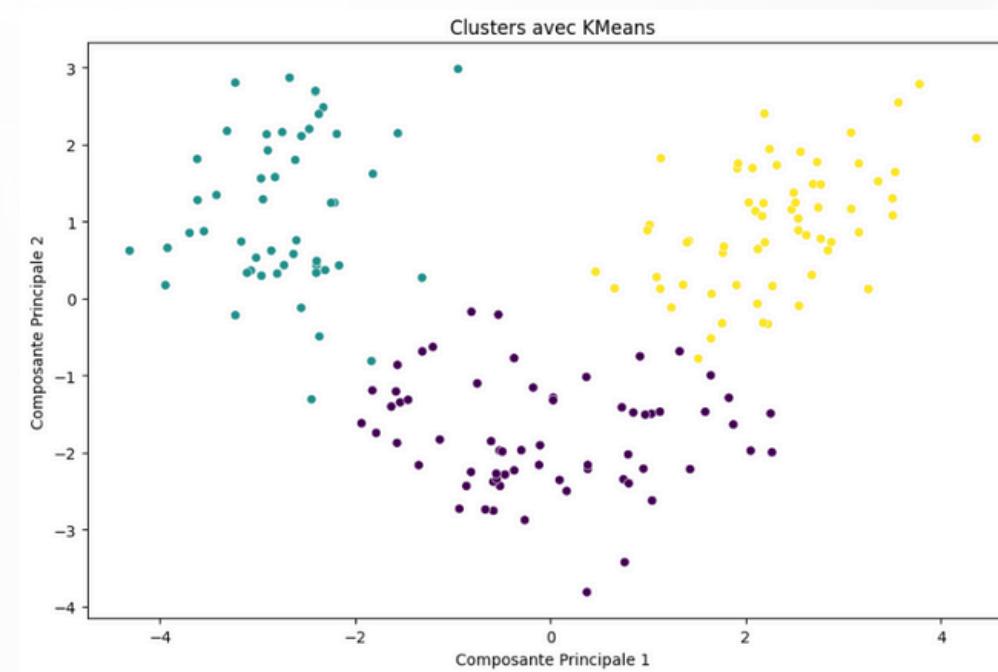
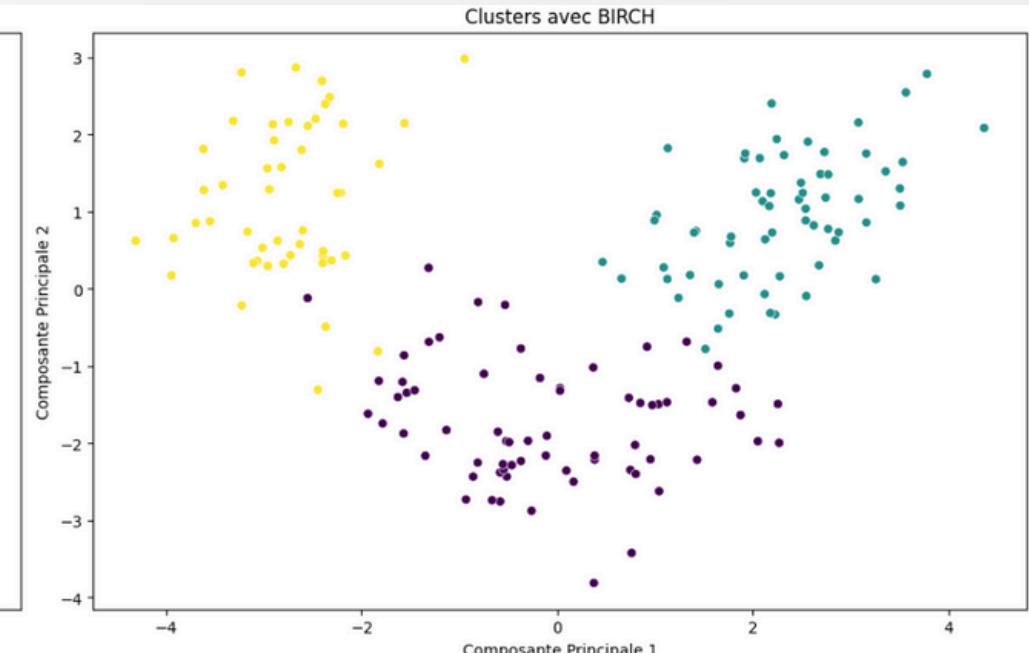
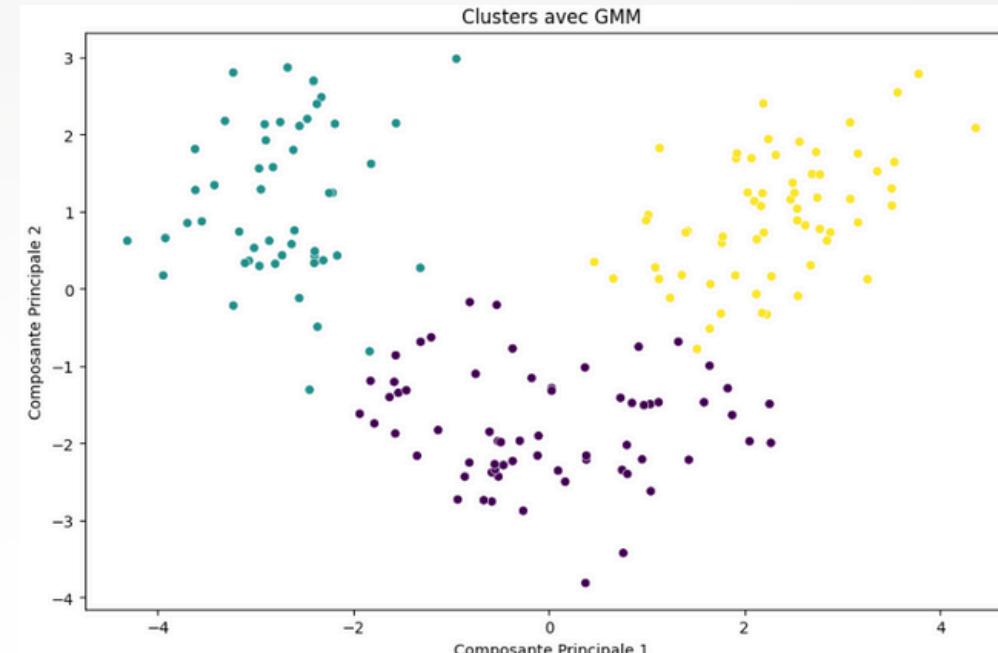
Les modèles ont été évalués selon trois métriques de clustering :

- **Silhouette Score** : Mesure la cohésion et la séparation des clusters. Plus la valeur est élevée, plus les clusters sont distincts.
- **Davies-Bouldin Score** : Indicateur de la dispersion des clusters, où une valeur plus faible est préférable.
- **Calinski-Harabasz Score** : Mesure la densité des clusters, où une valeur plus élevée indique des clusters plus denses et bien séparés.

ANALYSE DES RESULTATS

Comparaison des Performances

- Plusieurs algorithmes de clustering testés :
 - K-means, DBSCAN, CAH, GMM, BIRCH
- Évaluation des modèles avec :
 - Calinski-Harabasz Score
 - Coefficient de silhouette
- K-means :
 - Performances solides, scores élevés
 - Clusters denses et bien séparés
- DBSCAN :
 - Efficace pour détecter des clusters de formes variées
 - Limitations : tous les points classés comme bruit (100% noise ratio)
- GMM et CAH :
 - Résultats intéressants mais clusters moins compacts que K-means



ANALYSE DES RESULTATS

Choix du Modèle Final

- K-means sélectionné comme meilleur modèle
 - Simplicité, rapidité, et clusters bien définis
 - Capacité à segmenter en trois groupes distincts
 - Facilité d'interprétation pour analyse ultérieure
- Avantage par rapport à DBSCAN :
 - K-means crée des clusters compacts et significatifs
- Conclusion : K-means est retenu pour sa qualité de segmentation et sa simplicité d'interprétation

	Modèle	Silhouette Score	Davies-Bouldin Score	Calinski-Harabasz Score
0	KMeans	0.345719	1.151019	99.056198
1	CAH	0.340378	1.171293	96.991639
3	GMM	0.342062	1.164884	96.402000
4	BIRCH	0.338897	1.157332	96.026209

APPLICATION DÉPLOYÉE

L'application utilise le modèle K-means pour prédire le cluster d'un vin en fonction de ses caractéristiques chimiques, permettant une classification rapide et intuitive en saisissant simplement les valeurs des variables dans un formulaire.

Résultat de la Prédiction

Le vin appartient au cluster numéro :

2

[Faire une nouvelle prédiction](#)

Prédiction du Cluster de Vin

Alcohol	Malic Acid
<input type="text"/>	<input type="text"/>
Ash	Ash Alcancy
<input type="text"/>	<input type="text"/>
Magnesium	Total Phenols
<input type="text"/>	<input type="text"/>
Flavanoids	Nonflavanoid Phenols
<input type="text"/>	<input type="text"/>
Proanthocyanins	Color Intensity
<input type="text"/>	<input type="text"/>
Hue	OD280
<input type="text"/>	<input type="text"/>
Proline	
<input type="text"/>	

[Prédire](#)

LIMITES & AMÉLIORATIONS



Limites

- K-means : Sensible à l'initialisation et aux clusters non sphériques
- DBSCAN : Sensibilité aux paramètres (epsilon, MinPts) ; 100% des points classés comme bruit
- CAH et BIRCH : Scalabilité limitée ; difficulté à capturer des formes variées
- Données limitées : Taille de la base réduisant la représentativité des clusters

Améliorations:

- Prétraitement : Réduction du bruit, transformations non linéaires (e.g., Box-Cox)
- Optimisation : Recherche de paramètres optimaux (grid search, validation croisée)
- Algorithmes alternatifs : OPTICS, clustering par graphes, ICA + ACP
- Scalabilité : Algorithmes distribués (e.g., MiniBatch K-means) pour grands ensembles de données

**MERCI POUR VOTRE
ATTENTION**