Université Joseph Ky ZERBO — Burkina Faso

UFR : Sciences Informatiques Appliquées — La Patrie ou la Mort, nous Vaincrons

Filière : **Sciences des Données** — Année académique : 2024 – 2025

Niveau : **M1-S2** — Ouagadougou, le 13/11/2025

MACHINE LEARNING NON SUPERVISE

Projet : K-means et Interface Utilisateur : Analyse de la Qualité des Vins Blancs par Segmentation Non Supervisée

ENSEIGNANT : Dr Serge SONFACK SOUNCHIO

MEMBRES DU GROUPE 5 :

KAMBOULE Dagwankpouro

NAOUE Sampana

TOU M. A. Idrissa

OUENA Edouard

RAPPORT D'ANALYSE : SEGMENTATION K-MEANS DU DATASET *WINE QUALITY* (WHITE)

Ce rapport synthétise l'étude de segmentation non supervisée appliquée au jeu de données des vins blancs portugais (*Wine Quality – Vinho Verde*), en utilisant l'algorithme **K-means**. Le projet a été mené dans un cadre académique de Machine Learning Non Supervisé, en suivant une méthodologie rigoureuse allant de l'exploration des données au déploiement d'une application interactive (Streamlit).

1. Introduction et Objectifs du Projet

1.1. Contexte

La classification des vins est traditionnellement basée sur des critères sensoriels et géographiques. L'approche par *Machine Learning* non supervisé permet de découvrir des profils chimiques intrinsèques basés sur les données physico-chimiques, sans préjugé lié à une étiquette de qualité prédéfinie. Le K-means est l'algorithme de choix pour sa simplicité et son efficacité à identifier des groupes homogènes.

1.2. Objectifs

L'objectif principal du projet est de découvrir des clusters naturels au sein du dataset *Wine Quality* des vins blancs. Les buts secondaires sont :

- **Développer une chaîne de traitement** complète (prétraitement, standardisation, sélection de K) avec une approche basée sur des *bonnes pratiques* de Data Science.
- **Évaluer la qualité du clustering** par des métriques internes (Silhouette, DBI, CH) avec validation croisée.
- **Interpréter les clusters** en termes de caractéristiques physico-chimiques pour définir des **profils types** de vin.
- **Déployer une interface interactive (Streamlit)** permettant la configuration, l'entraînement et la prédiction.

1.3 structure du projet

```
├── wine-quality.ipynb        # Notebook d'analyse et de clustering K-means
├── streamlit_app.py          # Application Streamlit interactive
├── winequality-white.csv     # Dataset UCI (vins blancs, séparateur ;)
├── kmeans_model.joblib       # Modèle K-means sauvegardé
├── scaler.joblib             # StandardScaler sauvegardé
```

```
├── kmeans_model.pkl          # Sauvegarde alternative
├── scaler.pkl
├── model_metadata.pkl        # Métadonnées du modèle (K choisi, features, etc.)
└── README.md
```

Lien github: https://github.com/dkamboule/Jaccard_similarity.git
Lien du dataset: https://archive.ics.uci.edu/

1.4 Lancement de l'application

- Définition et activation de l'environnement de travail

```
python -m venv .venv
source .venv/bin/activate     # Linux/Mac
# ou
.venv\Scripts\activate      # Windows
```

- Mise à jour et installation des dépendances

```
pip install --upgrade pip

pip install -r requirements.txt
```

- Lancement de l'application streamlit

```
streamlit run streamlit_app.py
ou
python -m streamlit run streamlit_app.py
```

2. Données et Préparation

2.1. Le Jeu de Données

Le dataset winequality-white.csv comprend 4898 entrées décrivant des vins blancs par 11 caractéristiques physico-chimiques quantitatives.

Les variables clés incluent :

Variables	Description
fixed acidity	Acidité fixe
volatile acidity	Acidité volatile
residual sugar	Sucre résiduel
chlorides	chlorures
free sulfur dioxide	dioxyde de soufre libre
citric acid	acide citrique
total sulfur dioxide	dioxyde de soufre total
density	densité

pH	pH
sulphates	sulfates
alcohol	alcool

La variable quality (allant de 0 à 10) est utilisée uniquement pour l'évaluation a posteriori et non pour l'entraînement K-means.

2.2. Prétraitement et Normalisation

Une préparation rigoureuse est essentielle pour l'efficacité du K-means :

Etape de prétraitement	**Résultat**	**Justification**
Nettoyage des données	Suppression de 937 doublons, laissant 3 961 observations uniques	Assurer l'indépendance des observations.
Gestion des outliers	163 outliers détectés et potentiellement conservés	L'application Streamlit permet à l'utilisateur de choisir de les supprimer ou de les conserver.
Standardisation	Application du StandardScaler sur les 11 variables.	Obligatoire : K-means est sensible aux échelles. La standardisation centre les données et réduit l'impact des variables à forte variance.

3. Méthodologie et Modélisation K-means

3.1. Choix Optimal du Nombre de Clusters (K)

Le succès du K-means dépend du choix de K. Une approche multi-critères a été adoptée :

1. **Méthode du Coude (Elbow)** : Ne nous situe pas clairement sur la valeur de k à choisir.
2. **Scores de Qualité Interne** : Évaluation des Scores de **Silhouette**, **Davies-Bouldin (DBI)** et **Calinski-Harabasz (CH)**.
3. **Validation Croisée (K-Fold)** : Teste la stabilité des métriques pour différents K sur des sous-échantillons, renforçant la confiance dans le choix final.

Le consensus obtenu en croisant ces critères a désigné **K=2** comme la segmentation la plus stable et interprétable, malgré un score de Silhouette modéré (0.213), typique des données réelles.

3.2. Visualisation et Réduction de Dimension

Afin de visualiser la séparation des clusters, une Analyse en Composantes Principales (PCA) a été utilisée. La projection sur les deux premières composantes (PCA-2D) confirme visuellement la séparation des deux groupes identifiés par le modèle.

Métrique	**Valeur (K=2)**	**Interprétation**
Silhouette Score	0.213	Séparation des clusters modérée.
Davies-Bouldin Index (DBI)	1.797	Qualité de séparation acceptable.
Calinski-Harabasz Index (CH)	1006.735	Clusters relativement bien séparés et compacts.

4. Interprétation des Profils de Vins

L'analyse des centroïdes des deux clusters révèle des profils physico-chimiques bien distincts, particulièrement sur trois caractéristiques clés :

Caractéristique	**Cluster 0 (N=1925)**	**Cluster 1 (N=2036)**	**Caractère Dominant**
Alcool	11.17 %	9.61 %	Cluster 0 (Vins plus alcoolisés)
Sucre Résiduel	3.42 g/L	10.11 g/L	Cluster 1 (Vins plus doux/sucrés)
Dioxyde de Soufre Total	116.08 mg/L	172.70 mg/L	Cluster 1 (Vins avec plus de SO2)
Qualité Moyenne	6.04	5.54	Cluster 0 (Meilleure Qualité)

4.1. Profils Dégagés

- **Cluster 0 : Vins de Haute Teneur en Alcool** : Ces vins affichent un degré d'alcool significativement plus élevé, un faible sucre résiduel et une concentration en SO2 plus basse. Ce profil est associé à la meilleure **qualité moyenne** des deux groupes.
- **Cluster 1 : Vins Doux et Riches en SO2** : Ces vins sont caractérisés par une forte teneur en sucre résiduel et un niveau élevé de dioxyde de soufre total, associés à un degré d'alcool plus faible. Ce profil correspond à une **qualité moyenne inférieure**.

4.2. Outil de Déploiement

Le projet est déployé via une application **Streamlit (streamlit_app.py)** permettant à l'utilisateur de :

1. Configurer les hyperparamètres (choix de K, gestion des outliers).
2. Entraîner le modèle et visualiser les métriques en temps réel.
3. Interpréter les clusters via des graphiques radar (visualisation des profils normalisés).
4. Prédire l'appartenance à un cluster pour un nouveau vin saisi manuellement, offrant une valeur ajoutée opérationnelle.

5. Conclusion et Perspectives

5.1. Conclusion

L'implémentation du K-means a permis de segmenter efficacement le dataset des vins blancs en deux groupes distincts, basés principalement sur l'équilibre entre **alcool, sucre résiduel, et SO2**. Le projet a atteint ses objectifs de robustesse et de reproductibilité, fournissant un outil complet pour l'analyse et la prédiction.

5.2. Pistes d'Amélioration Futures

Les pistes d'évolution principales sont :

- **Tests Algorithmiques** : Examiner d'autres méthodes de clustering non-linéaires ou probabilistes (GMM, DBSCAN, MiniBatchKMeans).
- **Sélection d'Hyperparamètres** : Intégrer des techniques de sélection de features (RFE, Chi-carré) pour améliorer la pureté des clusters.
- **Étendue du Projet** : Intégrer l'analyse du dataset des vins rouges pour identifier des profils communs et spécifiques.
- **Déploiement Professionnel** : Conteneuriser l'application (Docker) pour faciliter l'intégration et le déploiement sur un serveur.