Rapport préliminaire de l'étude des vins *Vinho Verde*

CHABANCE Yann, FROMONT Pierre, LAPIE Joséphine

16/04/2025

Résumé

Dans le cadre de l'UV SY09 – *Science des données*, un projet est proposé aux étudiants, consistant en l'analyse approfondie d'un jeu de données selon une démarche scientifique. Ce document présente l'état d'avancement de notre projet à mi-parcours. Le rapport final complet sera rendu peu avant les soutenances.

Introduction

Ce document présente l'état d'avancement de notre projet réalisé dans le cadre de l'UV SY09 pour le semestre de printemps 2025.

La première partie (section 1) est consacrée à la description du jeu de données utilisé, ainsi qu'à ses principales caractéristiques. Dans un second temps, nous présenterons les premières analyses exploratoires menées, visant à approfondir la compréhension des données et des relations entre les variables (section 2). Enfin, le document se conclura par une liste d'axes d'amélioration et de pistes à explorer pour la suite du projet (section 3).

1 Présentation du jeu de données

Dans le contexte de ce projet, nous avons choisi d'utiliser le jeu de données *Wine Quality Dataset* (lien Kaggle). Celui-ci décrit différentes variantes des vins rouges *Vinho Verde* à travers 11 composés ou relevés chimiques, ainsi qu'une note de qualité allant de 0 à 10. Cette dernière variable entière résulte de la moyenne des avis de sommeliers. L'intégralité des vins est identifiée via un ID, représentant la treizième variable. Au total, 1143 individus sont exprimés en fonction de 13 variables quantitatives (Tableau 1).

Après une rapide analyse du *Wine Quality Dataset*, nous avons remarqué qu'aucun individu ne comportait des données manquantes. Cependant, 125 lignes du tableau étaient des copies parfaites d'autres individus. Nous avons décidé de les supprimer pour ne pas fausser les différentes analyses comme l'ACP.

Table 1 – Tableau des variables du jeu de données

Nom de la variable	Format
fixed acidity	float64
volatile acidity	float64
citric acid	float64
residual sugar	float64
chlorides	float64
free sulfur dioxyde	float64
total sulfur dioxyde	float64
density	float64
pH	float64
sulphates	float64
alcohol	float64
quality	int64
ID	int64

Finalement, nous nous retrouvons avec un tableau de 1018 vins prenant leur valeur dans :

- 11 variables quantitatives continues représentant les données chimiques des liquides.
- 1 variable quantitative discrète cible (*quality*)
- 1 variable quantitative discrète permettant l'identification des différents vins (*ID*)

2 Analyses exploratoires

2.1 Manipulations préliminaires

Avant toute analyse, il est important d'avoir un jeu de données cohérent pour obtenir des résultats satisfaisants. Il est donc nécessaire de vérifier, comme énoncé précédemment, si le *Wine Quality Dataset* présente des données vacantes ou des doublons. De plus, pour éviter les problèmes liés aux échelles des différentes variables, la normalisation est primordiale. En effet, les données *alcohol* ($\in [8.4, 14.9]$) pourraient avoir une trop grande influence sur les analyses comme l'ACP comparées aux données *sulphates* ($\in [0.33, 2]$).

Nous utilisons alors le `StandardScaler` de la librairie

`sklearn` appliquant le TCL aux données.

2.2 Corrélation

Une des premières analyses importantes lors de l'étude d'un jeu de données, est l'analyse de la corrélation entre les variables. En effet, cela permet de déceler des liens reliant les variables et de mieux comprendre les tendances et de simplifier les problématiques. Le package `seaborn` est composé de la fonction `heatmap` permettant justement de tracer un diagrame de corrélation (figure 1).

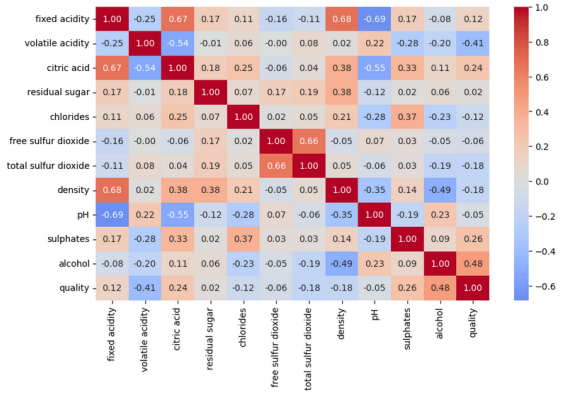

FIGURE 1 – Diagramme de corrélation des variables du dataset

Certaines corrélations intéressantes semblent se détacher de ce graphique et pourraient être utilisées :

- Corrélation inversée *pH* - *fixed acidity* : Logique vu que le pH est inversement proportionnel à l'acidité.
- Corrélation *quality* - *alcohol* : Très intéressante puisque la variable cible est assez corrélée à une variable en particulier

2.3 Analyse de la variable cible *quality*

La variable *quality* prend ses valeurs entre 0 et 10 (comme montré sur le tableau 2). Cependant, ces valeurs sont totalement déséquilibrées dans l'intervalle. En effet, comme elle représente une note subjective attribuée aux vins, elle est centrée autour de la note moyenne (5-6) et n'est jamais extrême. Plusieurs choix de modélisations sont donc possible :

TABLE 2 – Tableau des variables du jeu de données

quality	3	4	5	6	7	8
Nombre de vins	6	33	483	462	143	16

- *Modélisation à 6 classes* - La variable *quality* reste inchangée et les analyses supervisées donneront un résultat sur 10 entre 3 et 8.
- *Modélisation à 3 classes* - Modélisation via les quantiles empiriques des tiers donnant les classes 3-4-5, 6 et 7-8. Cela diminue les inégalités de distribution, mais aussi le sens de la classe *moyenne* qui correspond à la note 6/10.
- *Modélisation à 2 classes* - Les valeurs sont divisées en deux, en formant deux classes (*basse* et *haute*) équitable et ayant du sens. De plus, cela facilite l'analyse supervisée qui pourra être efficace.

Plusieurs stratégies ont été utilisées pour perfectionner ces modélisation dont la méthode du coude, le score de Silouhette ou la fonction de répartition empirique. Au final, la dernière option semble la plus adéquate.

2.4 ACP

L'Analyse en Composantes Principales finalisera notre rapport sur l'avancée à mi-parcourt du projet. Nous utilisons le `PCA` du `sklearn`. Grâce au diagramme des inerties expliquées cumulées de chaque composante principale, nous avons choisi de limiter cette analyse à 5 composantes, représentant pratiquement 80% de l'inertie expliquée (figure 2).

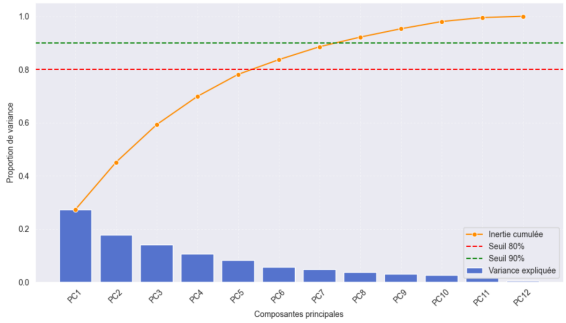

FIGURE 2 – Graphique des inerties expliquées cumulées

Le passage d'une analyse de 11 variable à seulement 5 permet de faciliter les prochaines étapes.

3 Avancées possibles

Plusieurs avancées sont prévues par la suite :

- Etude approfondie de l'ACP
- Clustering exploratoire
- Modélisation supervisée du dataset