Prédiction et Analyse de la Qualité des Jus Basée sur les Propriétés Chimiques

Youssef Hergli & Mohammed Ayoub Salma
Encadrante : Dr. Fatma Sbiaa

Année Universitaire 2024/2025

Contents

Abstract

Ce travail a pour objectif de développer un modèle d'intelligence artificielle (IA) permettant de classer la qualité du jus et de déterminer s'il est de bonne qualité, ainsi que d'estimer sa note de qualité. Ce projet vise à prédire précisément la qualité du jus à partir de ses propriétés mesurées, tout en optimisant la précision et la fiabilité du processus d'évaluation.
Mots-clés : Régression, RMSE, R^2, Random Forest, KMeans,PCA

1 Introduction

La qualité des produits alimentaires, notamment des jus, est un facteur clé pour garantir la satisfaction des consommateurs et le respect des normes de santé. L'évaluation de la qualité à partir d'analyses chimiques permet une estimation objective et automatisée, essentielle pour le contrôle industriel. Ce travail vise à développer un modèle prédictif efficace pour estimer la qualité des jus à partir de leurs propriétés chimiques détectées.

2 Travaux Liés

Des recherches antérieures ont exploré la prédiction de la qualité du jus à partir de propriétés chimiques :

- **Prédiction des attributs chimiques des agrumes à l'aide de réseaux de neurones artificiels et de régression linéaire** [1] : L'article rédigé par Mokhtar et al. (2016) porte sur l'évaluation de la qualité des fruits frais en se basant sur leurs caractéristiques chimiques. Cette étude compare l'efficacité de deux modèles prédictifs : le réseau de neurones et la régression linéaire. Les résultats montrent que le réseau de neurones offre une précision supérieure. Toutefois, d'autres méthodes comme les forêts aléatoires ou le clustering non supervisé n'ont pas été explorées.
- **Modeling wine preferences by data mining from physicochemical properties** [2] : Cet article de Cortez et al. propose de classifier la qualité des vins en fonction de mesures physico-chimiques. Plusieurs modèles de prédiction ont été utilisés, notamment la régression linéaire, les forêts aléatoires, les réseaux de neurones et les arbres de décision. La régression linéaire a montré les meilleures performances en termes de RMSE.

Les études de Mokhtar et al. (2016) et Cortez et al. (2009) montrent que l'efficacité des modèles varie selon la nature des données.

3 Méthodologie

3.1 Collecte et préparation des données

3.1.1 Dataset

Le jeu de données utilisé dans cette recherche pour le Sujet 6 a été fourni par Dr. Fatma Sbiaa. Il comprend 4 898 enregistrements correspondant aux tests de qualité du jus (lignes), chacun décrit par 12 variables représentant les caractéristiques chimiques

(colonnes), ainsi qu'une colonne supplémentaire indiquant la qualité du jus destinée à la classification.

Le dataset n'était pas parfaitement nettoyé. Nous avons donc appliqué plusieurs méthodes pour en faciliter l'exploitation :

- Suppression des valeurs manquantes ;
- Suppression des doublons.

3.1.2 Organisation des données

Le processus de préparation des données avait pour objectif de nettoyer et de formater l'ensemble des données en vue d'une analyse approfondie et d'une ingénierie des caractéristiques.

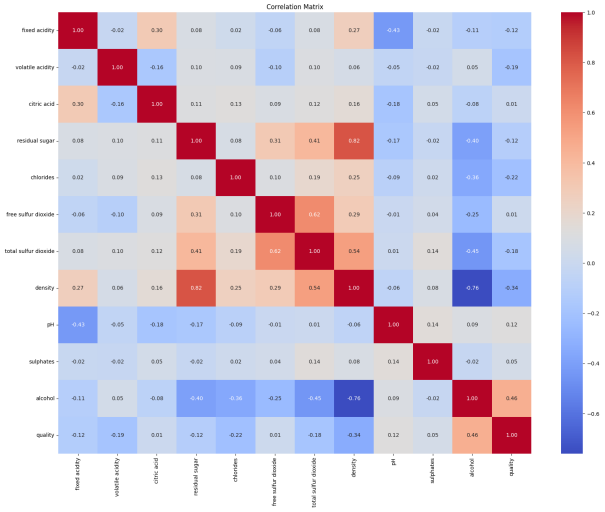

Figure 1: Matrice de corrélation entre les variables chimiques

Une analyse de l'importance des variables par rapport à la variable cible quality a été réalisée. Cette analyse permet d'identifier quelles caractéristiques chimiques influencent le plus la qualité du jus.

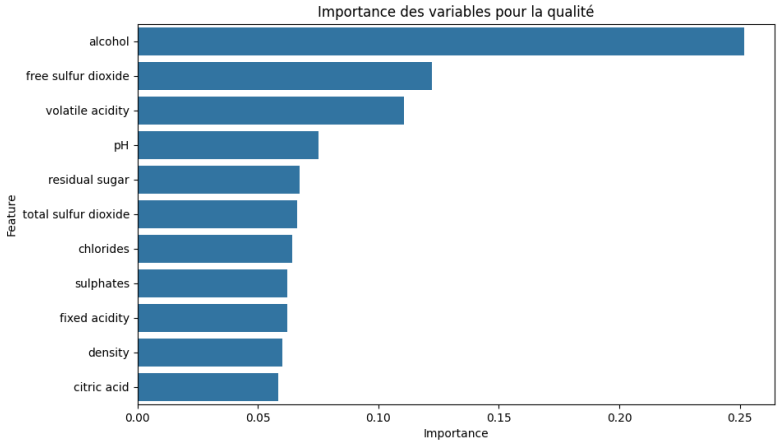

Figure 2: Importance des variables par rapport à la qualité du jus

La figure 2 montre que l'attribut *Alcohol* a la plus grande influence sur la qualité, suivi par le *free sulfur dioxide*. En revanche, des variables comme *density* et *citric acid* ont un impact relativement faible.

3.2 Construction des modèles :

On teste plusieurs modèles supervisés et non supervisés :

3.2.1 Modèles Supervisés

- *Régression Linéaire:*
  On utilise la régression linéaire : multiple, polynomiale et random forest regression pour donner une note de qualité au jus.
  Pour évaluer les performances des modèles de régression, deux métriques principales ont été utilisées :

  – **Coefficient de détermination (R^2)** :

  $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Plus R^2 est proche de 1, plus le modèle est meilleur.

  – **Root Mean Squared Error (RMSE)** :

  $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  Plus la valeur du RMSE est faible, meilleure est la qualité du modèle.

  Le test donne le résultat suivant :

Modèle	RMSE	R^2
Régression Linéaire Multiple	0.78	0.26
Régression Linéaire Polynomiale	0.84	0.14
Random Forest Regression	0.74	0.33

Table 1: Résultats des modèles de régression linéaire

  D'après le tableau 1, le Random Forest Regression a obtenu les meilleures performances, avec un RMSE de 0.74 et un R^2 de 0.33.
  Cela indique que ce modèle minimise le mieux l'erreur de prédiction et explique mieux la variabilité de la qualité du jus comparé aux modèles de régression linéaire simple et polynomiale.

- *Classification:*

Pour la classification, on utilise le modèle Random Forest Classifier.
Pour évaluer les performances du modèle de classification, plusieurs métriques sont utilisées

- **Précision (Precision)** :

$$\text{Précision} = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Positif}$$

- **Recall (Rappel)** :

$$\text{Recall} = \frac{Vrai\ Positif}{Vrai\ Positif + Faux\ Négatif}$$

- **F1-score** :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Recall}}{\text{Précision} + \text{Recall}}$$

Le test donne le résultat suivant :

Classe	Précision	Recall	F1-score	Support
3	0.00	0.00	0.00	5
4	0.43	0.10	0.16	30
5	0.56	0.56	0.56	234
6	0.52	0.70	0.60	349
7	0.49	0.27	0.35	145
8	0.25	0.03	0.06	30

Table 2: Résultats du modèle Random Forest Classifier

Le tableau 2 présente les métriques d'évaluation du modèle de classification. Le modèle Random Forest Classifier donne une précision globale de 53%, avec une meilleure capacité de rappel pour la classe 6. Cependant, la performance reste faible pour les classes minoritaires (3, 4 et 8), en raison du déséquilibre des données.

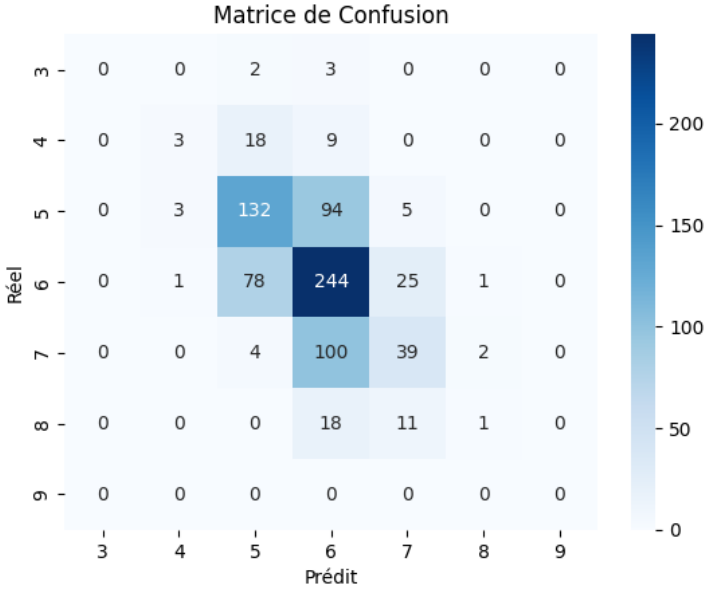

Figure 3: Matrice de confusion du modèle Random Forest Classifier

La figure 3 montre la matrice de confusion du modèle Random Forest Classifier. On observe que les prédictions sont majoritairement correctes pour les classes 5 et 6, tandis que les classes 3, 4 et 8 présentent davantage d'erreurs.

3.2.2 Modèles Non Supervisés (Clustering) :

Pour chercher le nombrre clustring exacte, on applique deux méthodes principales : Davies-Bouldin et la méthode de coude (kmeans)

- **KMeans** : La méthode du coude a été utilisée pour déterminer le nombre optimal de clusters.
  La figure suivante montre la courbe obtenue par la méthode du coude :

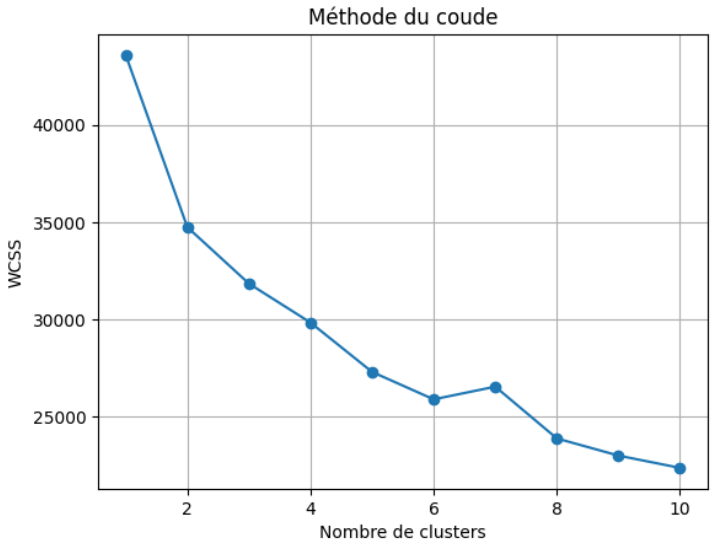

Figure 4: methode de coude

→ La methode du coude suggeére donne un nombre optimal de 3 ‘a 4 clusters.

- **Indice de Davies-Bouldin**:
  L'indice de Davies-Bouldin permet d'évaluer la qualité des clusters obtenus. Un indice plus bas indique un meilleur clustering.

La figure suivante montre l'évolution de l'indice de Davies-Bouldin :

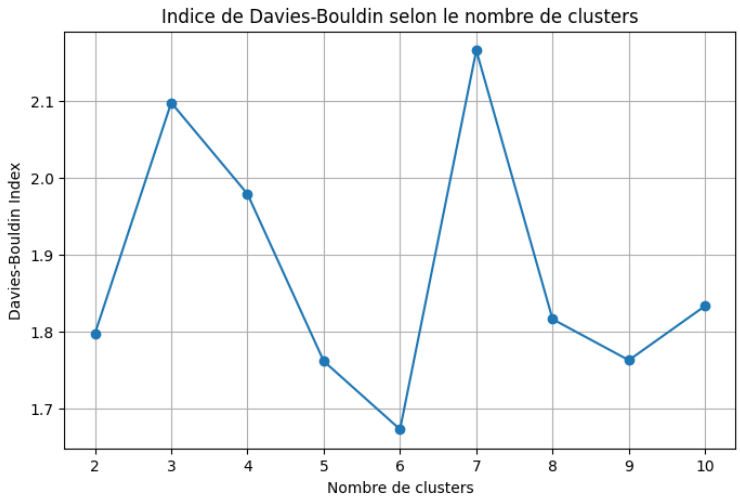

Figure 5: L'indice de Davies-Bouldi

L'indice de Davies-Bouldin atteint un minimum autour de 6 clusters.
→ d'aprés les deux analyses, un compromis a été choisi pour tester avec **3** & **6 clusters**.

- **Visualisation des clusters KMeans (k=3):**

  Après application de l'algorithme KMeans, les résultats ont été projetés en deux dimensions et trois dimensions en utilisant l'analyse en composantes principales (PCA).

  → **Visualisation en 2D :**
  La visualisation en 2D est présentée ci-dessous :

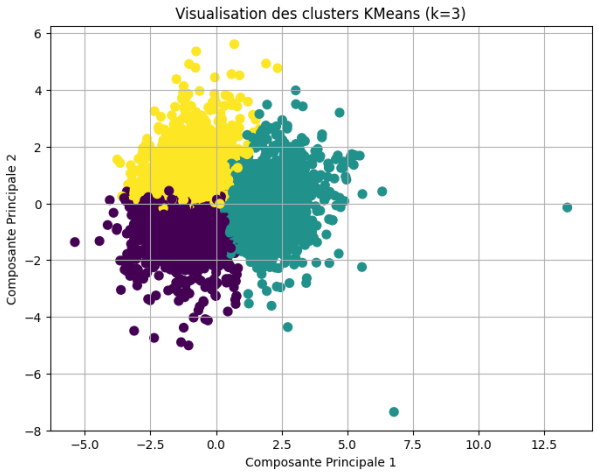

Figure 6: Visualisation des clusters KMeans (k=3) en 2D

→ **Visualisation en 3D :**

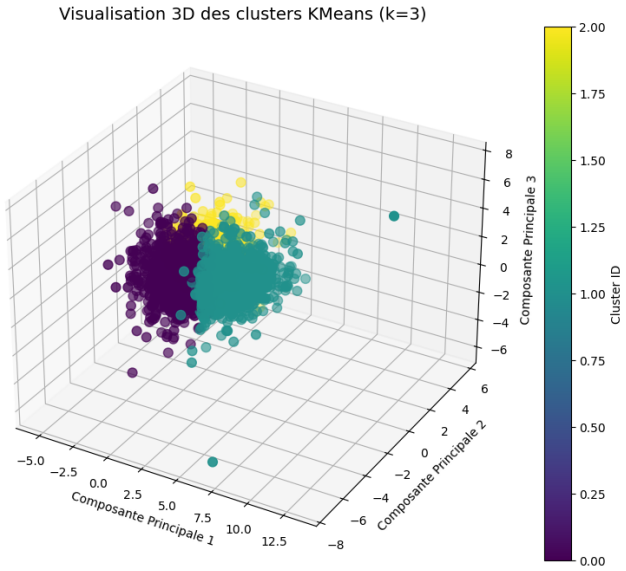

Figure 7: Visualisation des clusters KMeans (k=3) en 3D

→ La représentation en 3D confirme la bonne séparation de deux clusters, tandis qu'un chevauchement partiel est observable entre les deux autres.

- **Visualisation des Clusters KMeans (k=6):**

  → **Visualisation 2D :**

  La visualisation en 2D des clusters pour k = 6 est present ci-dessous :

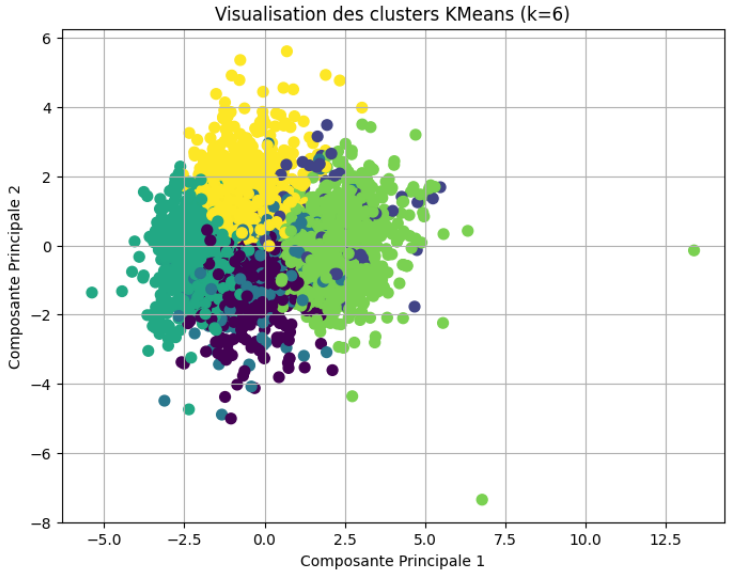

Figure 8: Visualisation des clusters KMeans (k=6) en 2D

→ **Visualisation en 3D :**

La visualisation en 3D des clusters pour $k = 6$ est présentée ci-dessous. Elle permet d'apprécier la répartition spatiale des différents groupes identifiés par l'algorithme KMeans.

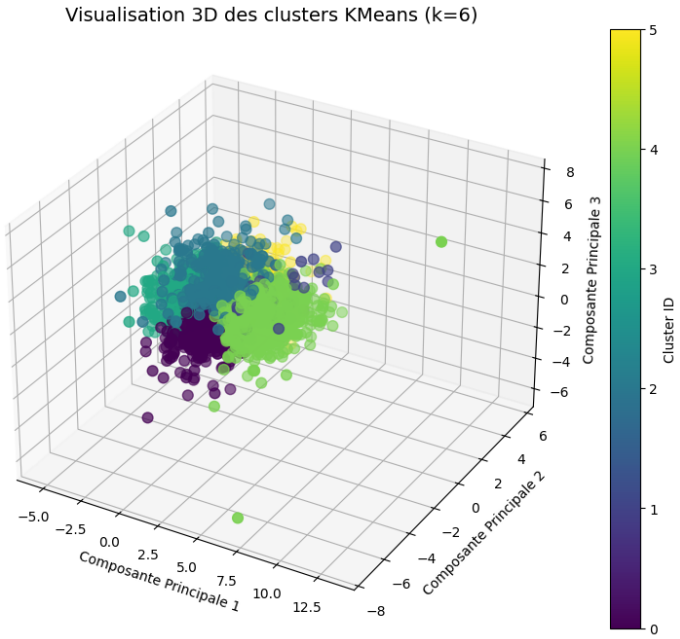

Figure 9: Visualisation 3D des clusters KMeans (k=6)

Plusieurs clusters sont fortement imbriqués dans l'espace tridimensionnel, indiquant que la séparation naturelle des données devient moins nette lorsque k augmente.

Afin de mieux comprendre la distribution des échantillons selon leur qualité, l'histogramme de la variable quality est analysé ci-dessous.

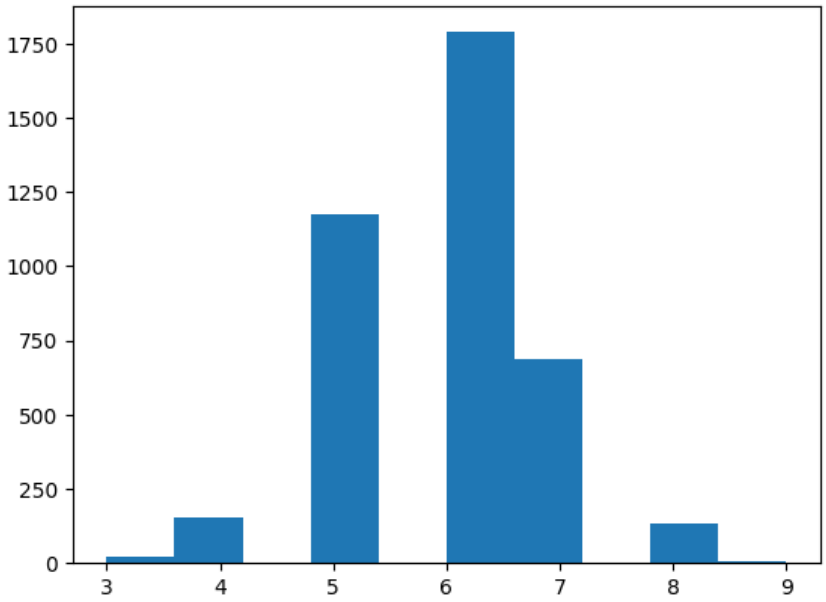

Figure 10: Histogramme de la variable qualité

L'analyse de la figure 10 révèle une forte concentration des tests autour des niveaux de qualité 5, 6 et 7, tandis que les notes extrêmes (3, 4, 8, et 9) sont très faiblement représentées.

Réorganisation des classes de qualité :
Pour obtenir une répartition plus équilibrée des classes et améliorer les performances des modèles de classification et de clustering, nous avons choisi de regrouper les qualités selon trois grandes catégories :

- **Basse** : qualité ≤ 5
- **Moyenne** : qualité $= 6$
- **Haute** : qualité ≥ 7

Cette nouvelle catégorisation permet de simplifier le problème de classification tout en respectant la distribution réelle des données.

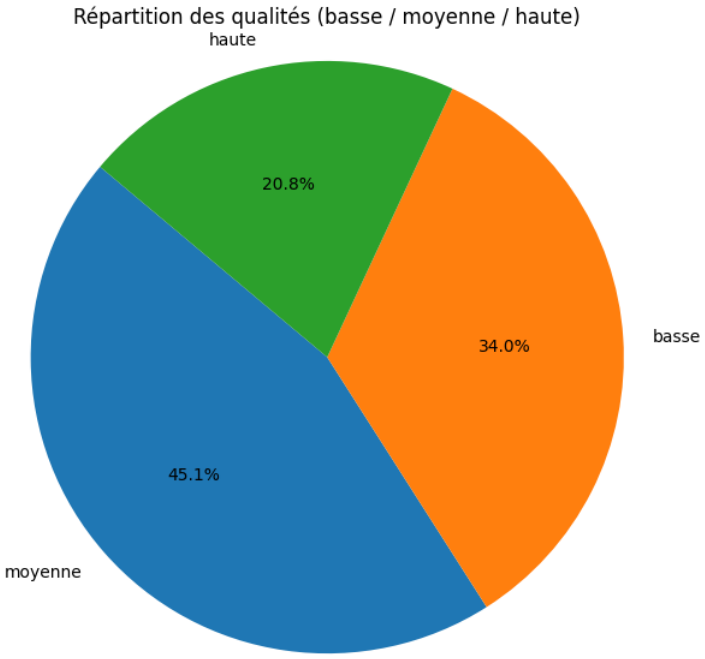

Figure 11: Répartition des qualités

4 Conclusion

Ce projet a permis de démontrer la faisabilité de la prédiction de la qualité des jus à partir de caractéristiques chimiques à l'aide de modèles de machine learning. La Random Forest Regression a offert les meilleurs résultats en termes de régression. Pour la classification, bien que le modèle Random Forest Classifier atteigne une précision globale de 53%, les performances restent limitées sur les classes minoritaires, mettant en évidence le besoin de traiter le déséquilibre des données. En clustering, les méthodes KMeans et l'indice de Davies-Bouldin ont montré que 3 ou 6 clusters sont des choix raisonnables, bien que la séparation soit parfois partielle. Ces résultats suggèrent qu'une approche plus sophistiquée combinant équilibrage des classes et optimisation des modèles pourrait encore améliorer les performances prédictives.

References

[1] Mokhtar, M., Selmi, T., & Ben Brahim, N. (2016). *Predicting Citrus Fruit Attributes Using Artificial Neural Networks and Linear Regression*. International Journal of Computer Applications, 139(4), 16-22. `https://www.mdpi.com/2311-7524/8/11/1016`

[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Modeling wine preferences by data mining from physicochemical properties*. Decision Support Systems, 47(4), 547-553. `https://doi.org/10.1016/j.dss.2009.05.016`