Prediction and Analysis of Juice Quality Based on Chemical Properties

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

Academic Year 2024/2025

Contents

Abstract

This work aims to develop an artificial intelligence (AI) model to classify the quality of juice and determine whether it is of good quality, as well as estimate its quality score. The project seeks to accurately predict juice quality based on its measured properties, while optimizing the accuracy and reliability of the evaluation process.
Keywords: Regression, RMSE, R^2, Random Forest, KMeans, PCA

1 Introduction

The quality of food products, particularly juices, is a key factor in ensuring consumer satisfaction and compliance with health standards. Evaluating quality based on chemical analyses allows for an objective and automated estimation, which is essential for industrial control. This work aims to develop an effective predictive model to estimate juice quality based on detected chemical properties.

2 Related Work

Previous research has explored predicting juice quality from chemical properties:

- **Predicting Citrus Fruit Attributes Using Artificial Neural Networks and Linear Regression** [1]: The article by Mokhtar et al. (2016) focuses on evaluating the quality of fresh fruits based on their chemical characteristics. This study compares the effectiveness of two predictive models: neural networks and linear regression. The results show that neural networks provide superior accuracy. However, other methods like random forests or unsupervised clustering were not explored.
- **Modeling Wine Preferences by Data Mining from Physicochemical Properties** [2]: This article by Cortez et al. proposes classifying wine quality based on physicochemical measurements. Several predictive models were used, including linear regression, random forests, neural networks, and decision trees. Linear regression showed the best performance in terms of RMSE.

The studies by Mokhtar et al. (2016) and Cortez et al. (2009) show that model performance varies depending on the nature of the data.

3 Methodology

3.1 Data Collection and Preparation

3.1.1 Dataset

The dataset used in this research for Topic 6 was provided by Dr. Fatma Sbiaa. It includes 4,898 records corresponding to juice quality tests (rows), each described by 12 variables representing chemical characteristics (columns), as well as an additional column indicating the juice quality intended for classification.

The dataset was not perfectly cleaned. We applied several methods to facilitate its exploitation:

- Removal of missing values;
- Removal of duplicates.

3.1.2 Data Organization

The data preparation process aimed to clean and format the dataset for in-depth analysis and feature engineering.

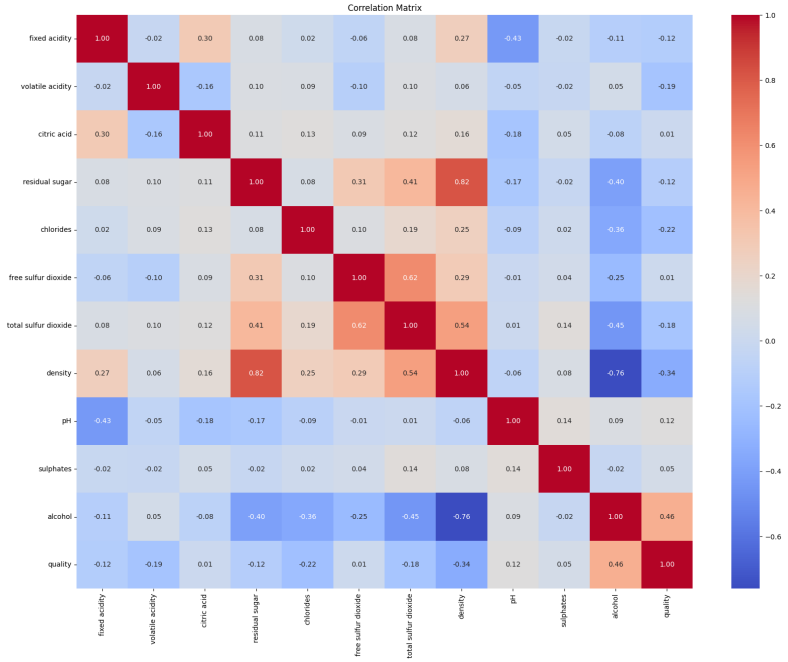

Figure 1: Correlation matrix between chemical variables

An analysis of variable importance with respect to the target variable *quality* was performed. This analysis identifies which chemical characteristics most influence juice quality.

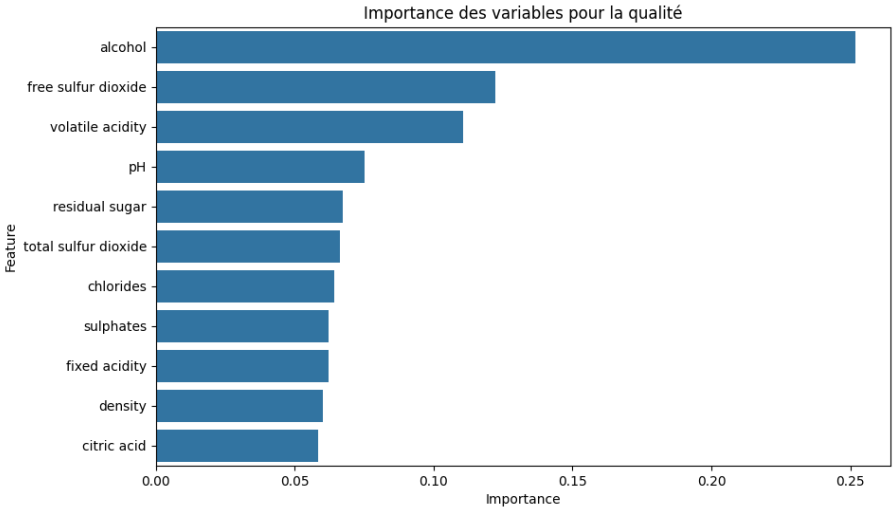

Figure 2: Importance of variables in predicting juice quality

Figure 2 shows that the attribute *Alcohol* has the greatest influence on quality, followed by *free sulfur dioxide*. In contrast, variables like *density* and *citric acid* have relatively low impact.

3.2 Model Construction

We tested several supervised and unsupervised models:

3.2.1 Supervised Models

- *Linear Regression:*
  We used multiple linear regression, polynomial regression, and random forest regression to assign a quality score to the juice.
  To evaluate the performance of regression models, two main metrics were used:
  - **Coefficient of Determination** (R^2):

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

    The closer R^2 is to 1, the better the model.
  - **Root Mean Squared Error (RMSE)**:

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

    The lower the RMSE value, the better the model's quality.

  The results are as follows:

Model	RMSE	R^2
Multiple Linear Regression	0.78	0.26
Polynomial Linear Regression	0.84	0.14
Random Forest Regression	0.74	0.33

Table 1: Results of linear regression models

According to Table 1, Random Forest Regression achieved the best performance, with an RMSE of 0.74 and R^2 of 0.33.
This indicates that this model minimizes prediction error and explains the variability in juice quality better compared to simple and polynomial linear regression models.

- *Classification:*

For classification, we used the Random Forest Classifier.
To evaluate the classification model's performance, several metrics were used:

- **Precision**:

$$\text{Precision} = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- **Recall**:

$$\text{Recall} = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- **F1-score**:

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

The results are as follows:

Class	Precision	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: Results of the Random Forest Classifier

Table 2 presents the evaluation metrics of the classification model. The Random Forest Classifier achieved an overall precision of 53%, with the best recall for class 6. However, performance remains low for minority classes (3, 4, and 8) due to data imbalance.

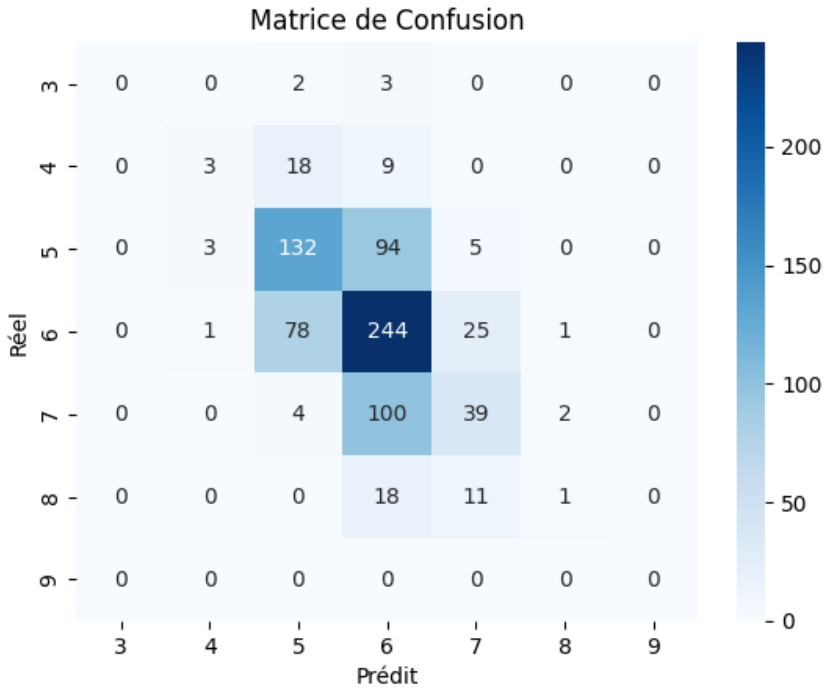

Figure 3: Confusion matrix of the Random Forest Classifier

Figure 3 shows the confusion matrix of the Random Forest Classifier. We observe that predictions are mostly correct for classes 5 and 6, while classes 3, 4, and 8 have more errors.

3.2.2 Unsupervised Models (Clustering)

To determine the optimal number of clusters, two main methods were applied: Davies-Bouldin index and the elbow method (KMeans).

- **KMeans**: The elbow method was used to determine the optimal number of clusters. The figure below shows the elbow curve:

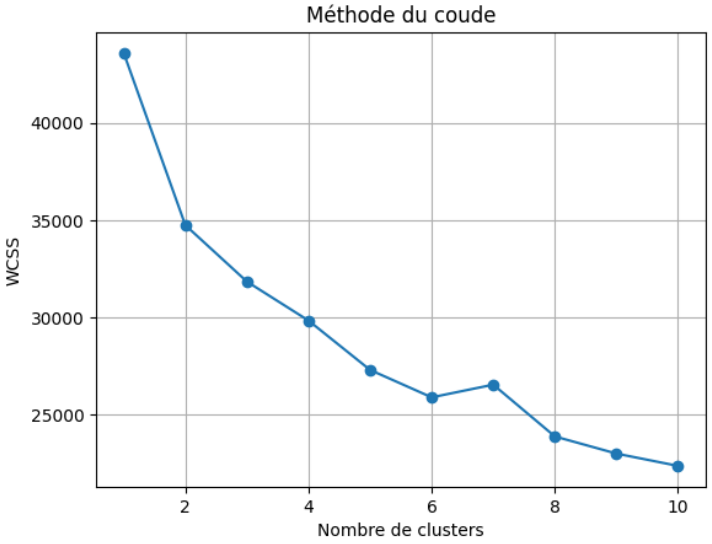

Figure 4: Elbow method

→ The elbow method suggests an optimal number of 3 to 4 clusters.

- **Davies-Bouldin Index**:
  The Davies-Bouldin index evaluates the quality of the obtained clusters. A lower index indicates better clustering.

The figure below shows the evolution of the Davies-Bouldin index:

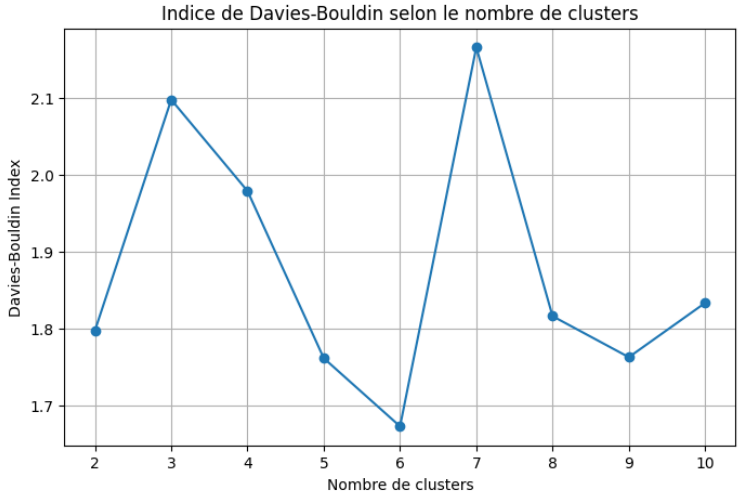

Figure 5: Davies-Bouldin Index

The Davies-Bouldin index reaches a minimum around 6 clusters.
→ According to both analyses, a compromise was made to test with **3** and **6 clusters**

- **Visualization of KMeans Clusters (k=3):**

  After applying the KMeans algorithm, the results were projected into two and three dimensions using Principal Component Analysis (PCA).

  → **2D Visualization:**
  The 2D visualization is presented below:

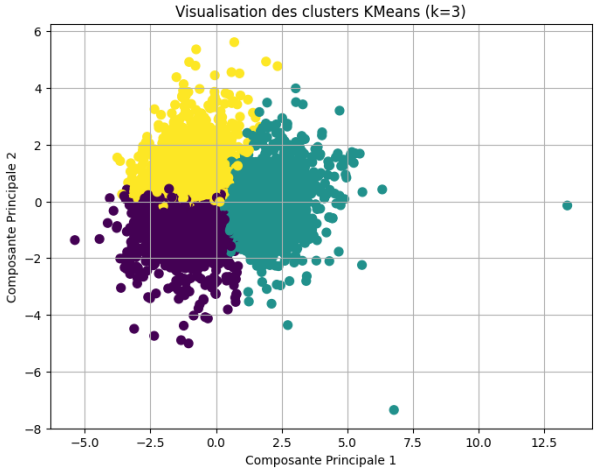

Figure 6: 2D visualization of KMeans clusters (k=3)

→ **3D Visualization:**

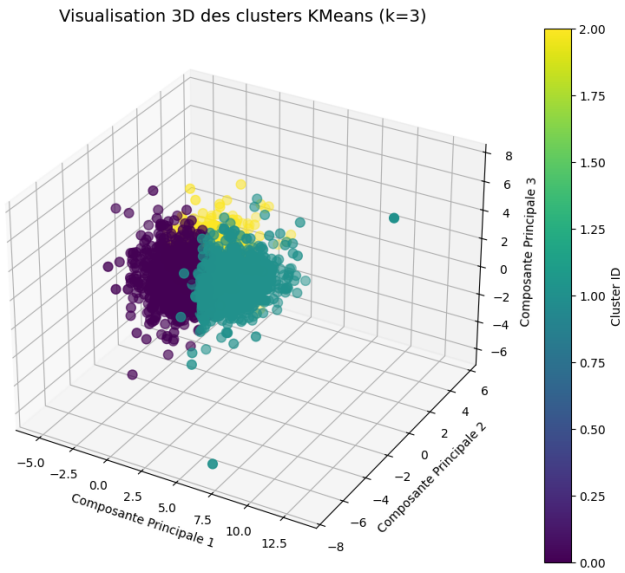

Figure 7: 3D visualization of KMeans clusters (k=3)

→ The 3D representation confirms a good separation between two clusters, while a partial overlap is observable between the other two.

- **Visualization of KMeans Clusters (k=6):**

  → **2D Visualization:**

  The 2D visualization of the clusters for $k = 6$ is shown below:

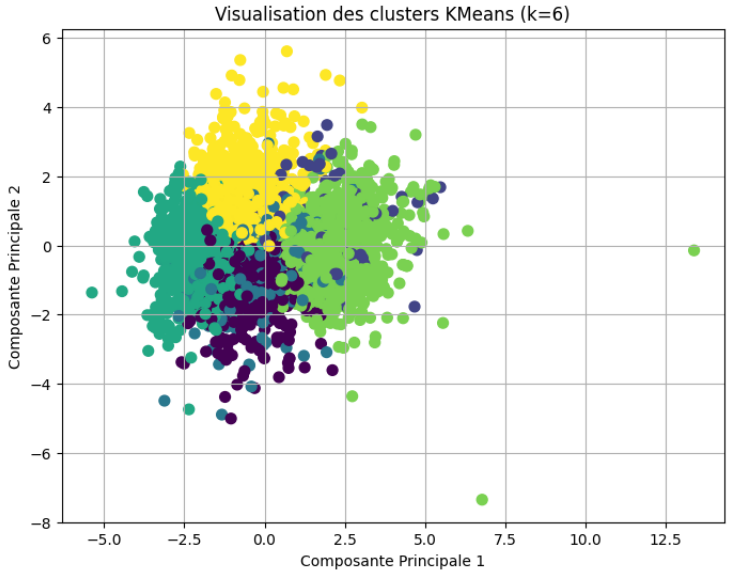

Figure 8: 2D visualization of KMeans clusters (k=6)

→ 3D Visualization:

The 3D visualization of the clusters for $k = 6$ is shown below. It illustrates the spatial distribution of the different groups identified by the KMeans algorithm.

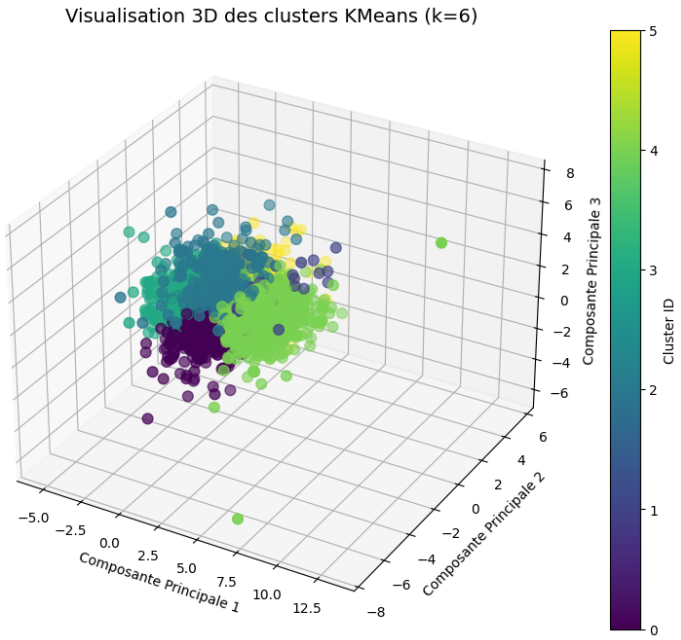

Figure 9: 3D visualization of KMeans clusters (k=6)

Several clusters are heavily intertwined in three-dimensional space, indicating that the natural separation of the data becomes less distinct as k increases.

To better understand the distribution of samples according to their quality, the histogram of the *quality* variable was analyzed:

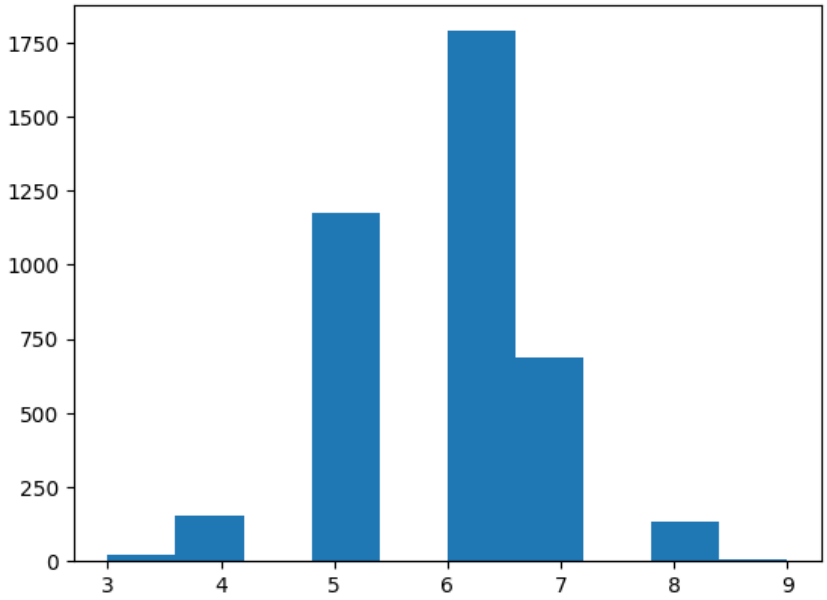

Figure 10: Histogram of the quality variable

The analysis of Figure 10 reveals a strong concentration of tests around quality levels 5, 6, and 7, while extreme scores (3, 4, 8, and 9) are very underrepresented.

Reorganization of Quality Classes:
To achieve a more balanced distribution of classes and improve the performance of classification and clustering models, we chose to regroup the qualities into three broad categories:

- **Low**: quality ≤ 5
- **Medium**: quality $= 6$
- **High**: quality ≥ 7

This new categorization simplifies the classification problem while respecting the real distribution of the data.

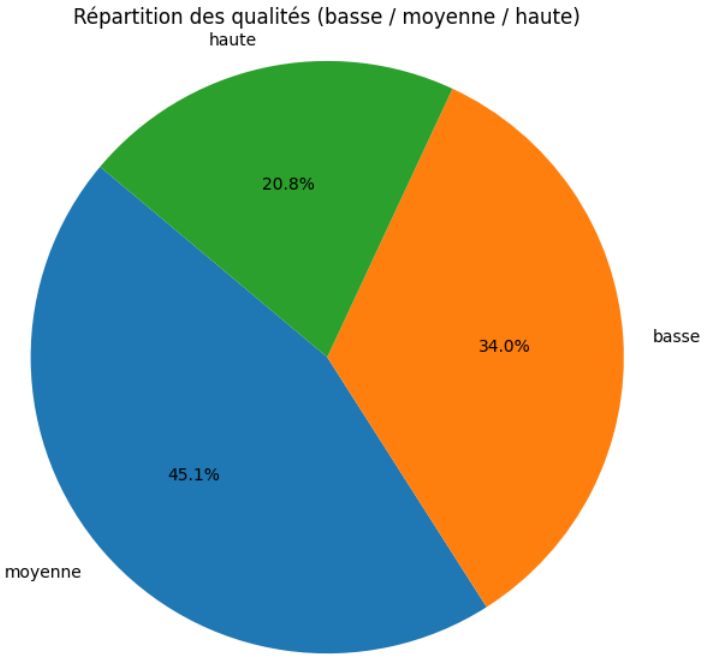

Figure 11: Distribution of quality categories

4 Conclusion

This project demonstrated the feasibility of predicting juice quality from chemical characteristics using machine learning models. Random Forest Regression provided the best results for regression tasks. For classification, although the Random Forest Classifier achieved an overall precision of 53%, performance remains limited on minority classes, highlighting the need to address data imbalance. In clustering, the KMeans method and Davies-Bouldin index indicated that 3 or 6 clusters are reasonable choices, although the separation is sometimes partial. These results suggest that a more sophisticated approach combining class balancing and model optimization could further improve predictive performance.

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`