Winery Analysis to Predict Type of Wine Through Multiple Models

By: Jenpicha Jenlarpwattanakul, Kalim Park, and Mikaella Valero
Group: Partners in Wine

Abstract:

The wine prediction data analyzes wine types and attributes based on the chemical compositions present during winemaking. Through the methods of linear regression, K means clustering/PCA, and random forest, the goal of identifying the patterns and groupings in the characteristics of wine was successfully performed.

I. Introduction

The production of wine presents an intricate and careful procedure to ensure proper bottling for consumers. Wine is created through the steps of grape growing, harvesting, pressing, fermentation, filtration, aging, and bottling. Understanding each and everyone of these procedures is important for maintaining consistency and quality per bottle. Beyond the physical process of wine production, the chemical components affect the overall taste for each wine. Factors such as alcohols, sugars, organic acids, pigments, and more all influence how wine develops its sensory characteristics.

Predicting wine can be a challenging task due to the variety of factors that affect the richness and appearance of wine. This paper explores the application of chemistry of winemaking and data science by finding the relationships of physico-chemical data through K-means clustering, linear regression, and random forest analysis. The selection of our data set presents the chemical and physical properties of wine, along with an accessible tabular data for our models and predictions to be created.

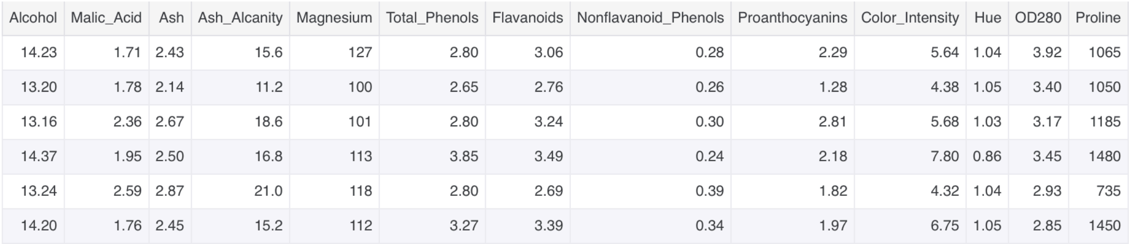
II. Literature Review

Whether it would be red wine, white wine, or rosé (pink wine), the cultivation of wine is dependent on the chemical composition. The presentation of wine is entirely reliant on the type of grape used, with red wines using deeper red grapes, white wines using white grapes, and pinker wines using only reminiscence of red grapes. Each grape has its own chemical compounds, minerals, and vitamins, which impacts the wine quality [3]. Machine learning has been an important tool as the center of technology grows, therefore impacting the development of machine models to improve accuracy and prediction. Because the physico-chemical of wine consists of numerous factors, machine learning can aid in resolving that and condensing it into specific features to be studied [3]. According to the UC Irvine wine dataset, in which our dataset was derived from, a featured technique of Principal Component Analysis (PCA) can be applied with dimensionality reduction in order to truly observe proper K-means patterns [1]**.** To improve the certainty and accuracy of wine predictions, coding a Random Forest model can aid in the data to make more informed decisions. Random Forests are fundamental in machine learning because of its ability to segment the data through a decision tree and create a final decision [4]. Based on these reviews towards similar wine prediction analyses, we believe that applying K-means clustering with PCA, linear regression, and random forest to be able to contribute towards the prediction of the wine being used for our specific data set.

III. Dataset Description:

The data set utilizes the models of K-means clustering and Linear Regression to only optimize the number of clusters and their visualization. The goal of our input towards this dataset is to make a prediction on what type of wine was used based on the characteristics presented above. The data was derived from Kaggle by the creator Xavier Vivancos García [2]. A notable suggestion that was done in the code was the removal of the column "Customer_Segment" due to the unnecessary data for wine classification.

- Wine.csv: This data set is comprised of 3 unknown types of wine that contains 13 characteristics of: *Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavonoids, Non Flavonoid Phenols, Proanthocyanidins, Color Intensity, Hue, OD280, and Proline*. This dataset has 145,750 interaction



Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Phenols	Flavanoids	Nonflavanoid_Phenols	Proanthocyanins	Color_Intensity	Hue	OD280	Proline
14.23	1.71	2.43	15.6	127	2.80	3.06	0.28	2.29	5.64	1.04	3.92	1065
13.20	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.40	1050
13.16	2.36	2.67	18.6	101	2.80	3.24	0.30	2.81	5.68	1.03	3.17	1185
14.37	1.95	2.50	16.8	113	3.85	3.49	0.24	2.18	7.80	0.86	3.45	1480
13.24	2.59	2.87	21.0	118	2.80	2.69	0.39	1.82	4.32	1.04	2.93	735
14.20	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450

Fig 1: Wine data based on the first 7 rows and 13 columns.

IV. Exploratory Data Analysis:

The 13 attributes of Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavonoids, Non Flavonoid Phenols, Proanthocyanidins, Color Intensity, Hue, OD280, and Proline were analyzed to explore the relationship between any of them. Figure 2 and figure 3 explains that out of all of the 13 attributes, phenol and flavonoids correlate the most with each other. The closer it is to 1, the more closely correlated they are, while the shades of blue indicate a more negative correlation. It points out that Total phenols and flavonoids have a high correlation coefficient of 0.87, indicating a strong positive relationship. Wines with higher total phenols are likely to have more flavonoids, given their dependent relationship.

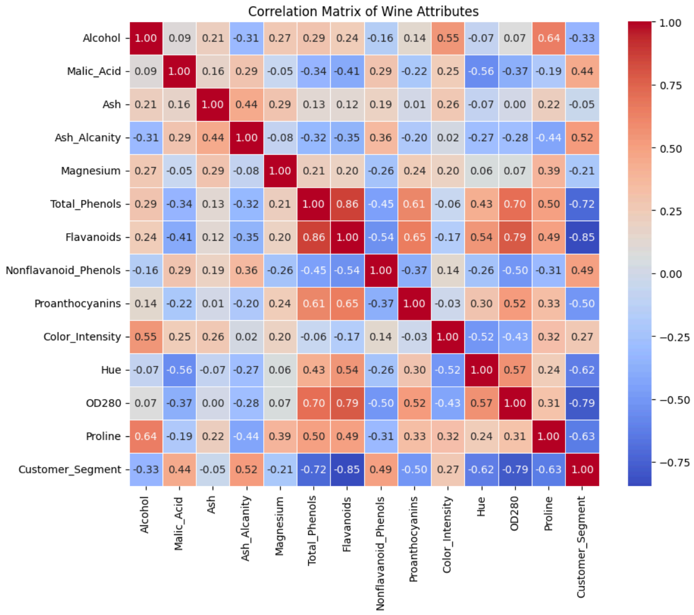


Fig 2: Heat map correlation matrix of Wine attributes.

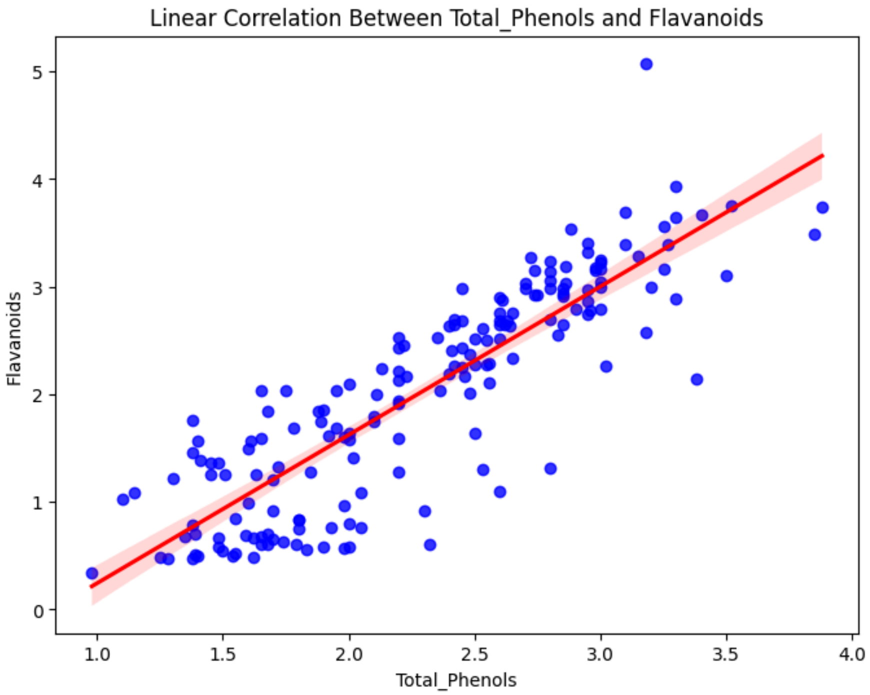


Fig 3: Linear correlation between total phenols and flavonoids.

V. Model Evaluation

A. Histogram

The histogram represents the value of each attribute including the x-axis represents attributes such as alcohol and the y-axis represents the count. To analyze this histogram, the height of each bar represents how it often happens and the each shape of histogram indicates overall range of values for each attribute

B. Linear Regression

The Linear Regression is showing the relationship between chemical components of wine. Between the total phenols and flavonoids, R^2 = 0.86 which represents the strong correlations. This tells that a significant portion of the variability in flavonoid could be explained by total phenols.

C. K-Means Clustering

The K-Means clustering process involves clustering data based on all columns, determining K clusters for all points, reducing dimensionality to two using PCA, plotting the first two principal components, and coloring the points according to their assigned clusters. We've done code using K=3, indicating three clusters with colored clusters such as yellow, blue and purple.

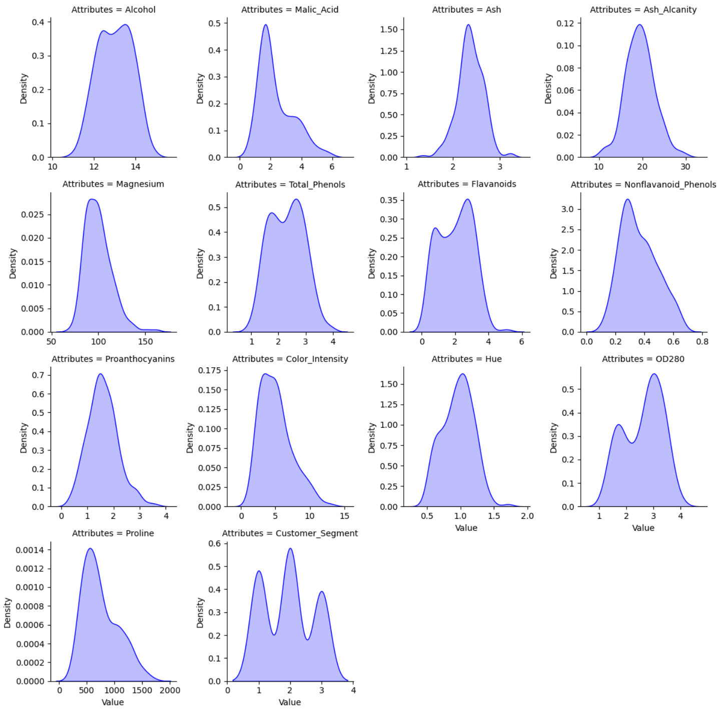
D. Principal Component Analysis (PCA)

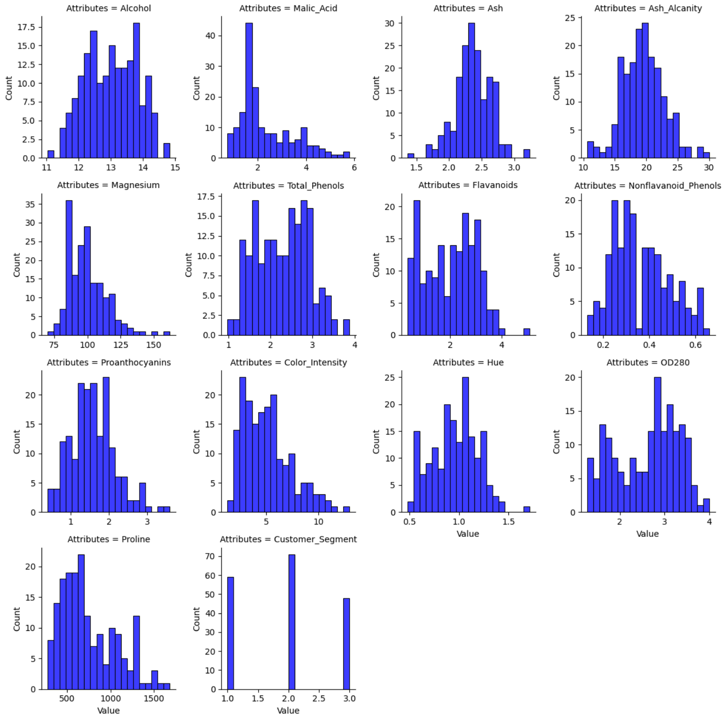
PCA is a method for retaining information by reducing the numbers in a dataset. Through this method, we applied PCA for 13 features and used the first two PCA.

E. Random Forest

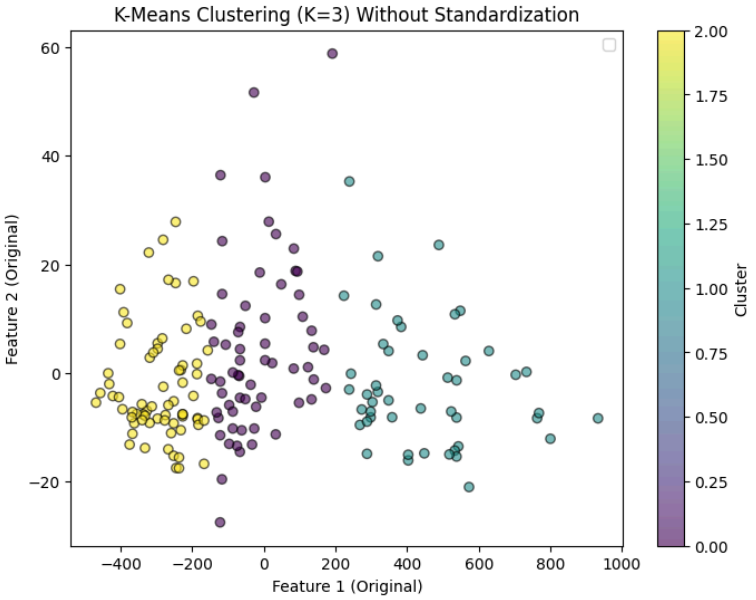
The Random Forest has been used to improve accuracy. It has achieved 100% of perfect accuracy in predicting all classes with no misclassifications of the confusion matrix showing no off-diagonal values which indicates perfect classification. Through the random forest classification report, we can see that all precision, recall and f1-score have 1.00 which is reflecting good model performance. In addition, since the macro and weighted average are 1.00 which indicates consistent performance across all classes.

VI. Visualization:





The histogram shows the distribution of attributes related to wine. Each graph shows the values of an attribute distributed across different counts in the dataset. In this histogram, the x-axis represents the attribute values and the y- axis represents the frequency of occurrence of values.



VII. Result & Interpretation

There is a strong correlation between phenol and flavonoids.

The main characteristics of the yellow cluster are a moderate alcohol level(12.17-13.11), low malic acid (0.94-1.45), varying flavonoids (0.57-3.18), non-flavonoid phenols(0.19-0.53), varying color intensity(1.95-5.75), balanced Hue(1.05-1.45), and moderate Proline(335-520). Based on these results, this cluster likely represents medium-bodied red wine such as Merlot and Grenache.

The key characteristics of the purple cluster are a high alcohol level(12.93-14.06), medium to high acidity (1.63-3.80), moderate to high flavonoid content (2.41-3.17), balanced non-flavonoid phenols (0.17-0.39), varying color (3.52-5.65), hue (0.96-1.12), and high proline content (735-845). Based on these characteristics, this cluster suggests full-bodied red wine such as Cabernet Sauvignon and Syrah/Shiraz.

The primary characteristics of the blue cluster are lower alcohol(12.17-13.11), low acidity(0.94-1.45), varying flavonoids (0.57-3.18), low non-flavonoid phenols(0.19-0.53), varying color intensity(1.95-5.75), and moderate proline levels (355-520). Overall, this cluster represents light-bodied wines such as Pinot Noir and Gamay.

VIII. Conclusion & Discussion

The wine prediction data aims to analyze wine attributes and types based on the chemical properties of winemaking. Through the correlation matrix and linear regression methods, there is a strong correlation between total phenol and flavonoids. K-means clustering/PCA show three clusters, each one representing a different type of red wine: medium-bodied red wine, full-bodied red wine, and light-bodied wine.

References

1. Aeberhard, S. & Forina, M. (1992). Wine [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.
2. Garcia, Xavier Vivancos (2020). Tutorial: Clustering wines with k-means. https://www.kaggle.com/code/xvivancos/tutorial-clustering-wines-with-k-means/input
3. Jain, K., Kaushik, K., Gupta, S. K., Mahajan, S., & Kadry, S. (2023). Machine learning-based predictive modelling for the enhancement of wine quality. *Scientific reports*, *13*(1), 17042. https://doi.org/10.1038/s41598-023-44111-9
4. Piyush Bhardwaj, Parul Tiwari, et al., A machine learning application in wine quality prediction. Machine Learning with Applications. Volume 8. 2022. 100261. ISSN 2666-8270. https://doi.org/10.1016/j.mlwa.2022.100261.