Winery Analysis

ANS 128
Team Name: Partners in Wine

Mikaella Valero, Kalim Park, Jenpicha Jenlarpwattanakul

Abstract

Summary

The wine prediction data analyzes wine attributes based on the chemical properties of 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 and led to a closer prediction of the wine used for the study.

Key findings

- The models revealed strong correlations between certain wine attributes such as phenols and flavonoids
- Accuracy - Random Forest method achieved a very good accuracy in predicting the wine.
- Clustering - K-means clustering utilizing three clusters to identify alcohol content, acidity, and tannin levels, which correspond to different wine characteristics.

Introduction

Background

Producing wine is an intricate process, with multiple steps to ensure proper fermentation and aging of the wine.
The assessments of wine are based on the factors of acidity levels, alcohol concentration, color intensity, etc.

Key Research Objective

The study aims to create models that analyze the relationship between the wine attributes (alcohol, malic acid, phenols, etc) and apply that to what type of wine it can be

Relevance

This is relevant in the Agricultural Science Industry because it applies to viticulture. Winemakers can utilize the chemical and physical analyses of their grapes/ wines and be able to produce consistent and profitable wine for their consumers.

Literature Review

Source: UC Irvine Machine Learning Repository

- Dataset includes 13 attributes used in wine classification and analysis.
- Past studies: Classification, clustering, and regression to predict wine quality, identify wine types, and determine optimal wine production techniques.
- Past method: Studies used Random Forest, Support Vector Machines(SVM), and K-Nearest Neighbors(KNN) which showed a similar perfect accuracy prediction
- How this project contributes
    - This project contributes by improving the prediction of the wine type based on the chemical properties through machine learnings such as Linear Regression, K-means, Correlations. It also aims to provide more accurate classifications towards quality of the wine.

Dataset Description

Source

- UC Irvine Machine Learning Repository

Key features

- Analysis determined the quantities of 13 attributes found in each of the three types of wines.
- Results of a chemical analysis of wines grown in the same region in Italy but derived from three different varieties.

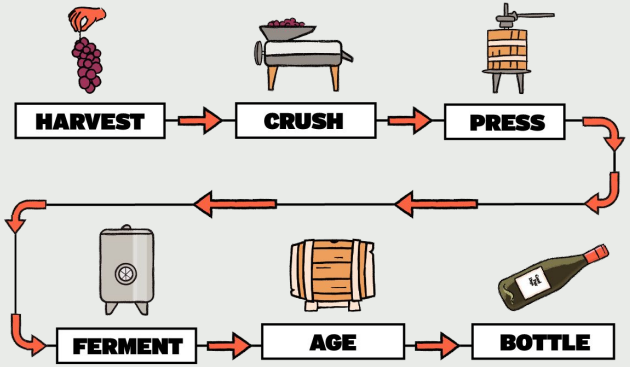
Data Preprocessing Step

- Initial data set had around 30 variables, but the owner lost some and only have the 13 dimensional version.

Dataset Description

The 13 chemical attributes are:

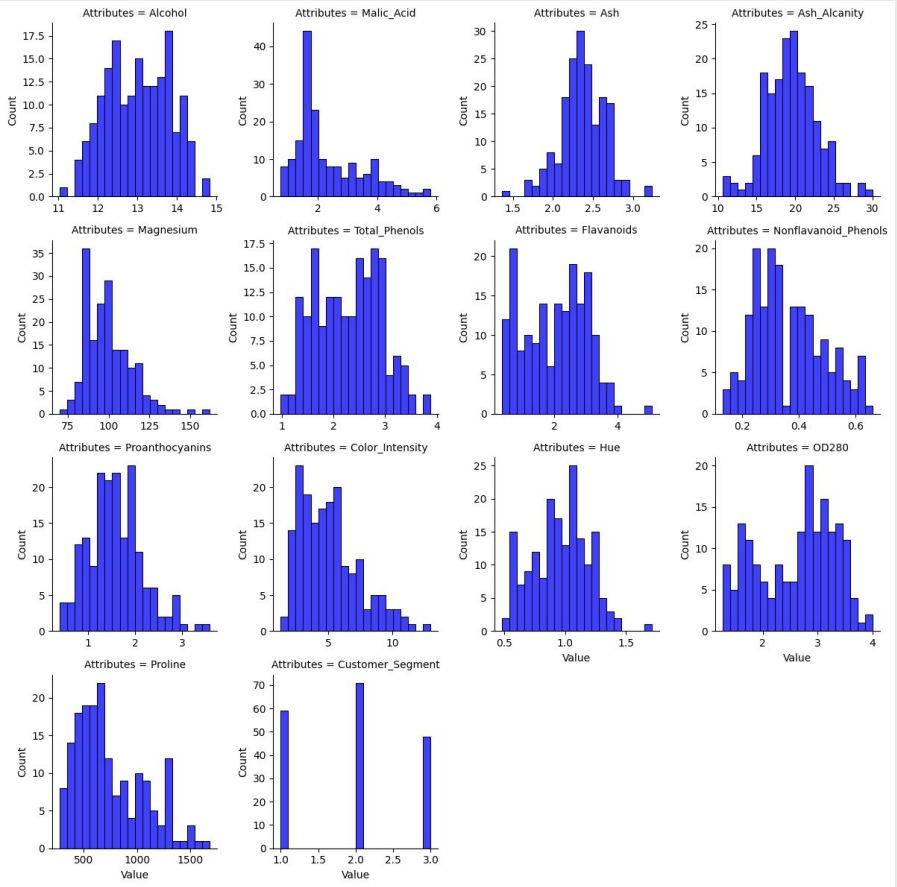
1. Alcohol: Helps the wine ferment and age over time
2. Malic Acid: affects tartness
3. Ash: neutralizes acidity
4. Ash Alkalinity: neutralizes acidity
5. Magnesium: aids in yeast metabolism and growth
6. Total Phenols: contribute to flavor and color
7. Flavonoids: antioxidant
8. Non Flavonoid Phenols: impacts taste
9. Proanthocyanidins: impacts color, flavor, astringency
10. Color Intensity: proportional to its flavor (deep colors = full bodied)
11. Hue: indicates wines age and grape variety
12. OD280: protein concentration
13. Proline: amino acid that helps mouthfeel of wine



Visualization

Histogram

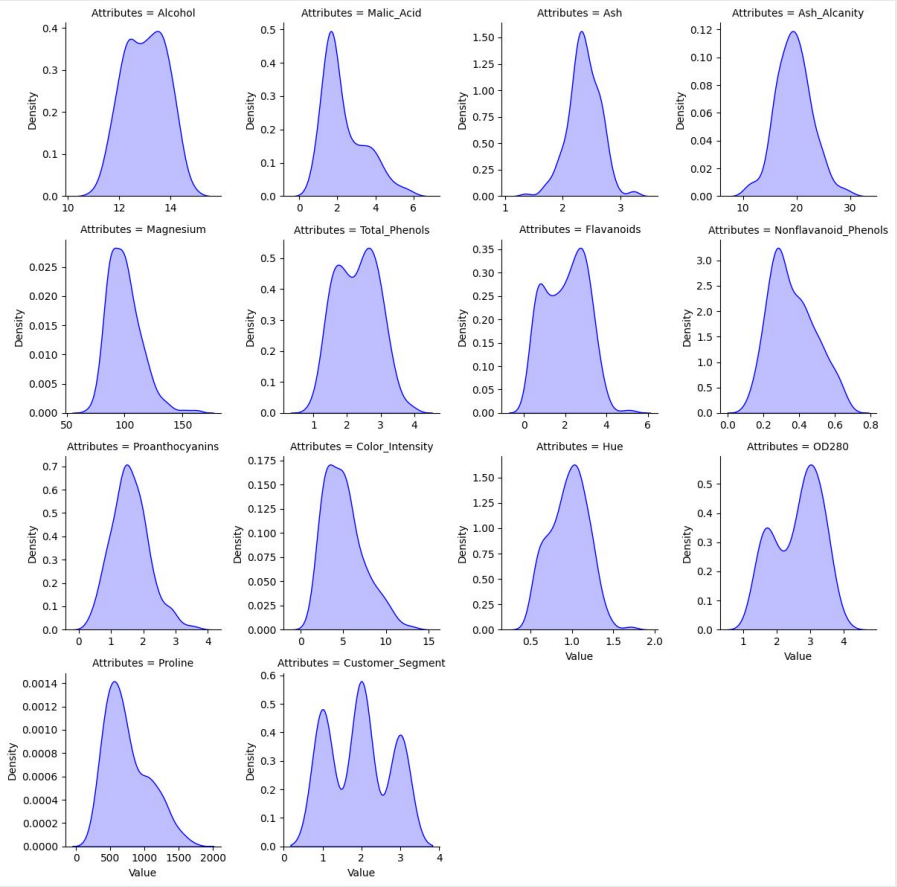
- How the values each attribute are distributed
- Helps identify the spread and skewness of data.
- Highlights potential outliers in the dataset.



Visualization

Density Plot

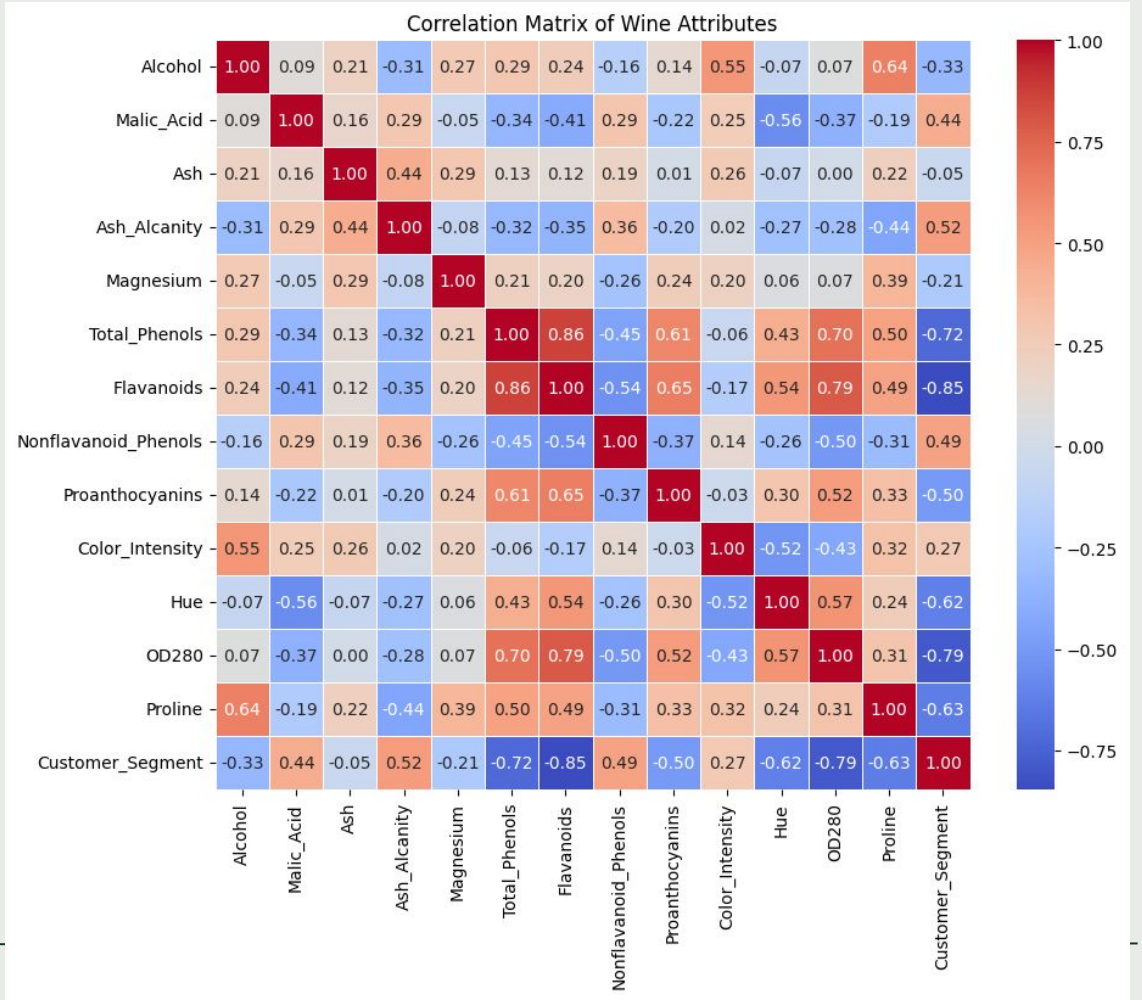
- Clearer accurate graph of the data's overall shape.
- Concentration of data points within ranges.



Exploratory Data Analysis

Correlation Matrix

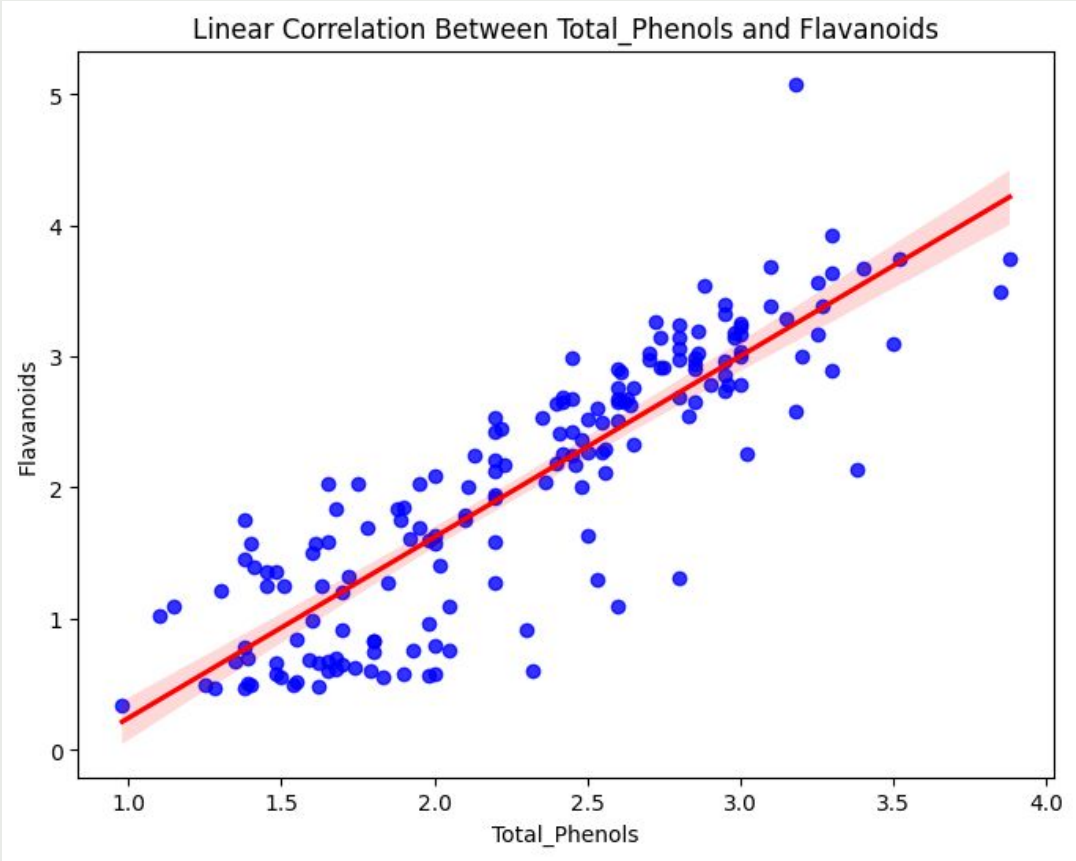
0.86 : correlation in total phenols and flavonoids



Exploratory Data Analysis

Linear Correlation Between Total Phenols And Flavonoids

R2 = 0.86

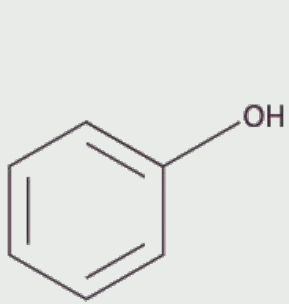


Phenol

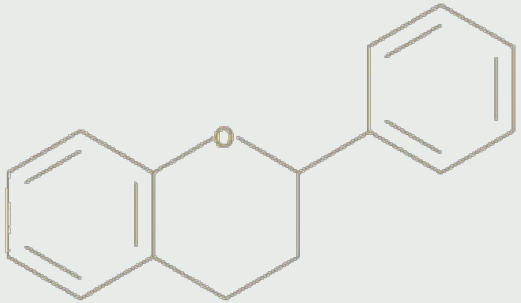
- Also known as polyphenols
- Natural compounds found in wine that affect its color, taste, and mouthfeel.
- Derived from the grape berry, especially the skins and seeds.

Flavonoids

- A class of phenols
- They are derived from the skin, seeds, and stems of grapes

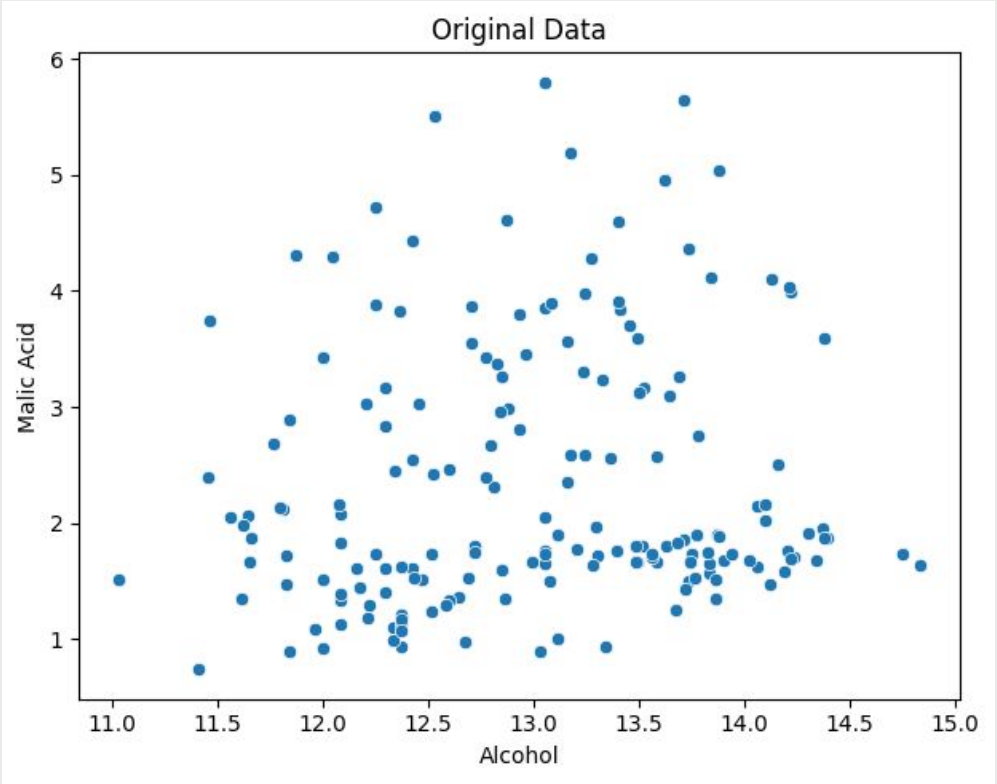


Phenol



Flavanoid

Scatter Plot
Alcohol vs. Malic Acid



- No strong relationship between Alcohol and Malic Acid.
- The data points are spread with no clear trend.
- Weak correlation observed (0.09).

Random Forest

- Perfect Accuracy
  - Model achieved 100% accuracy in predicting all classes.
- No Misclassifications
  - The confusion matrix shows no off-diagonal values, indicating perfect classification.
- Precision, Recall, F1-Score have 1.00, reflecting good model performance.
- Both macro and weighted average are 1.00, indicating consistent performance across all classes.

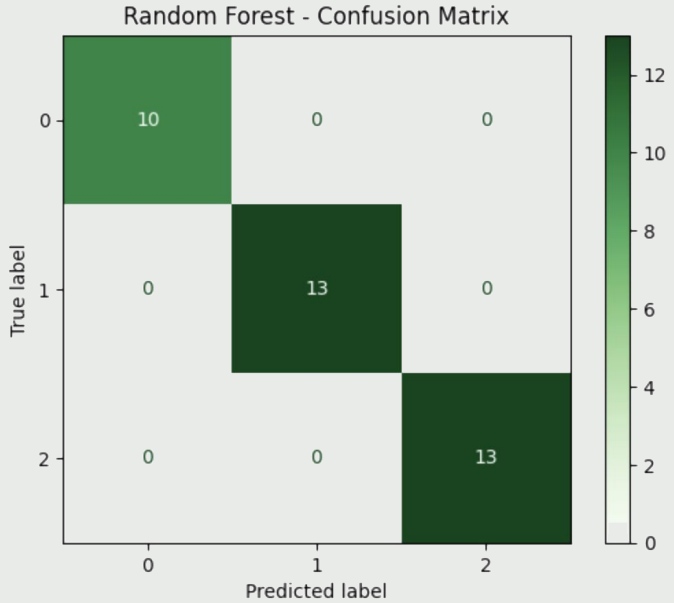
```
Random Forest Accuracy: 1.00

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        10
           1       1.00      1.00      1.00        13
           2       1.00      1.00      1.00        13

    accuracy                           1.00        36
   macro avg       1.00      1.00      1.00        36
weighted avg       1.00      1.00      1.00        36

<Figure size 600x400 with 0 Axes>
```

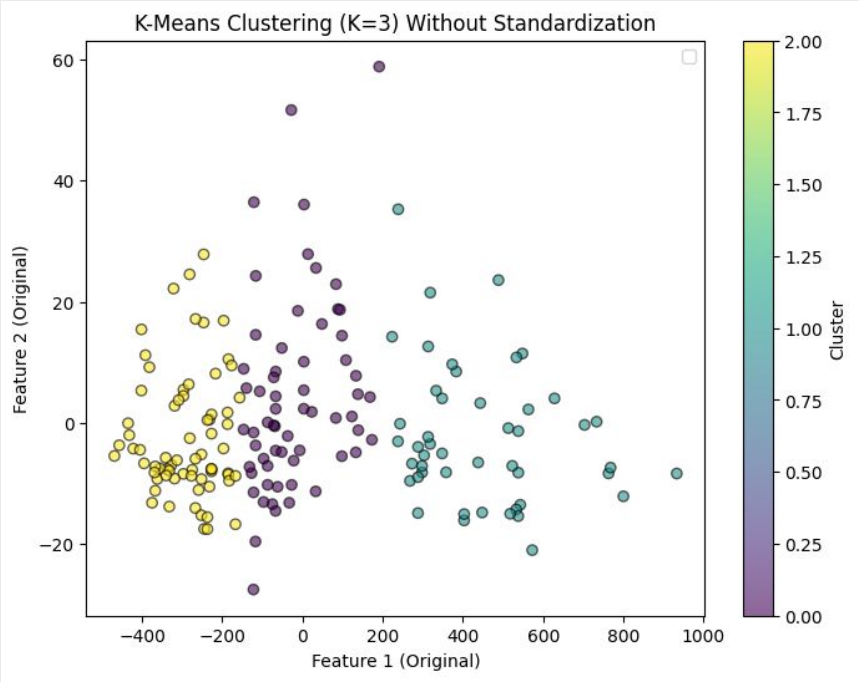


Steps to achieve K-mean Clustering

1. Cluster by all columns
2. Get K clusters for all points
3. Reduced data to 2 dimensions by PCA
4. Plot first 2 PC
5. Color points in the plot by the K clusters

PCA: Principal Component Analysis

K-Means Clustering (K=3) Without Standardization



- Data divided into 3 clusters (yellow, purple, and teal).
- Cluster Distribution: The plot displays how data points are grouped based on PC

Result

Yellow Cluster	Purple Cluster	Blue Cluster
• Moderate alcohol wines with lower acidity. • Diverse flavonoid levels, meaning some wines might be more bitter or astringent while others are smoother. • Varied color intensity, from light to deep tones. • Balanced proline levels, indicating medium-bodied wines.	• High alcohol • Medium-to-high acidity wines. • Rich in tannins and phenolic compounds, indicating structure and potential for aging. • Deeply colored wines, • High Proline levels, meaning full-bodied wines with a ripe fruit profile.	• Lower alcohol • Low acidity • Less intense in tannins, less bitter • Lighter to medium color intensity • Smooth, balanced sensation
Most likely : Medium-Bodied Red Wine - Merlot - Grenache	Most likely : full-bodied red wines - Cabernet Sauvignon	Most likely : light to medium-bodied wines - Pinot Noir - Gamay

Conclusion

- Using techniques such as Linear Regression, Random Forest, and K-means clustering, achieved high accuracy in classifying wines and identifying distinct clusters based on factors.
- The Random Forest method achieved perfect classification accuracy, providing a reliable approach for wine prediction.
- Performance of the Analysis successfully demonstrated predicting wine types by analyzing key chemical attributes.

Reference

https://archive.ics.uci.edu/dataset/109/wine