Section	Gloria - s232437	Mikolaj - s232534	Pier - s242943
1	40%	30%	30%
2	30%	40%	30%
3	30%	30%	40%
4	Equal	Equal	Equal

Table 1: Group contribution percentages per section

02450 - Introduction to Machine Learning and Data Mining

Project 01

Author

Gloria Stucchi - s232437 Mikolaj Jochim - s232534 Pier Franesco Sorgiovanni - s242943

October 3, 2024

Contents

1 Data Description

1.1 Problem Statement

The Wine dataset from the UCI Machine Learning Repository is used extensively in classification and clustering research, focusing on the problem of predicting the cultivar of wine based on its chemical properties. This dataset contains 178 instances of wines grown in the same Italian region but derived from three different cultivars.
Each instance is characterized by 13 numerical attributes, including alcohol content, malic acid levels, ash content, flavonoids, phenols, and other significant compounds that define the wine's quality. The overall problem of interest is to determine which chemical attributes most effectively distinguish between the different cultivars.

1.2 Data Reference

The data was obtained from the UCI Machine Learning Repository. The full reference is as follows: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. `https://archive.ics.uci.edu/dataset/109/wine`. Irvine, CA: University of California, School of Information and Computer Science.

1.3 Previous Data Analysis

1.3.1 Paper Title: *"Comparative analysis of statistical pattern recognition methods in high dimensional settings" [1]*

In this study, researchers used the wine recognition dataset containing 13 chemical attributes to evaluate various classification methods. The dataset's clear structure and adequate sample size provide a robust basis for testing classification efficacy.

The tested methods included:

- **Parametric Methods** (QDA and LDA): These assume specific data distributions, with QDA modeling distinct class covariances and LDA using a common covariance matrix for all classes.
- **Regularized Discriminant Analysis (RDA)**: This method extends QDA by regularizing covariance matrices to balance bias and variance, ideal for high-dimensional settings.
- **K-Nearest Neighbors (KNN)**: As a non-parametric method, KNN classifies based on proximity to nearest neighbors but struggles in high-dimensional spaces.

Other methods, such as Fisher's Discriminant Plane, were used for visualization and comparative analysis. In terms of results, RDA outshone all with a 100% Probability of Correct Classification, perfectly differentiating between classes. QDA and LDA also demonstrated robust performance, while KNN and two-stage methods exhibited limitations in high-dimensional contexts.

The study highlights RDA's effectiveness in environments where managing model complexity and avoiding overfitting are crucial, recommending it for datasets where feature dimensionality significantly affects class differentiation.

1.3.2 Paper Title: *"DE_PSO_SVM: An Alternative Wine Classification Method Based on Machine Learning "* [2]

This study has used the UCI Wine dataset with all 13 attributes to propose a novel classification approach for wine quality. Through the use of data enhancement, particle swarm optimisation to find optimal parameters for Gaussian kernel functions to be passed into a support vector machine for actual classification. The study also used K-nearest-neighbour, random forest and classification regression tree to cross test wine classification. The results found that their coined "DE_PSO _SVM" classification model returned the best results on all three wine models, and on the UCI Wine dataset had an **average precision of 0.997**, **average recall of 0.998** and an **average F1-score of 0.997**.

What they did to the data:

The authors of the paper performed a statistical analysis that included finding the mean, standard deviation, minimum, maximum and the quartiles of the 13 attributes. This was done to investigate whether normalisation of the data was required. Correlation of all 13 attributes was then computed using a visual matrix as a way to determine which were the most significant positive correlations to reduce the features dimensions, and hence reduce the complexity of calculations. It was also determined that there was to few samples within the Wine dataset to train a high accuracy model, and hence the data was enhanced by generating additional synthetic training samples based on the original 178.

1.4 Reflections on Classification and Regression

In the context of analysing wine data, the goal is understand the relationships between the various chemical components of the different wines to be able to perform classification techniques with an objective to categorise what region the tested wine has come from. The wines attributes will be strongly indicative of the area/cultivar, such as proline content which has correlations to the grapes growing temperature and the cultivars timing of harvest. Other attributes that may help in classification include flavonoids which are linked to the type of grape and the soil, while colour intensity again is also ifluenced by the climate and the winemaking techniques employed which can lead to the statistical identification of cultivars.

By using regression techniques, the goal is to predict continuous attributes. We aim to predict alcohol content or color intensity. This type of regression analysis will allow us to model the relationships between multiple chemical properties and the target variable which for example is alcohol, which is crucial for understanding factors that influence wine's flavor, quality and alcohol level both for safety and taxation. Relevant attributes to predict the alcohol content may include; proline which is correlated to grape ripeness which impact alcohol content as ripe grapes have more sugars, colour intensity which is linked to grape type where some grapes may have higher content of sugars.

To prepare the data for regression and classification we will require some transformations. First and foremost normalisation of the data will be necessary as the proline attribute values are significantly larger than the other attribute values and hence to ensure that all features contribute to the model equally we would need to normalise. Additionally to prevent potential over fitting problems with the model, we will need reduce some dimensions where certain attributes might be highly correlated this can either be done by performing a correlation matrix and removing one of the two significantly correlated values or using principle component analysis.

2 A detailed explanation of the attributes of the data

2.1 Attributes Categorization

Table 2: Table containing all attributes in the UCI dataset and the respective data type of that attribute

Attribute	Type of Data
alcohol	Continuous, Ratio
malic acid	Continuous, Ratio
ash	Continuous, Ratio
alcalinity of ash	Continuous, Ratio
magnesium	Continuous, Ratio
total phenols	Continuous, Ratio
flavonoids	Continuous, Ratio
nonflavonoid phenols	Continuous, Ratio
proanthocyanins	Continuous, Ratio
color intensity	Continuous, Ratio
hue	Continuous, Ratio
OD280/OD315 of diluted wines	Continuous, Ratio
proline	Continuous, Ratio
region	Discrete, Nominal

2.2 Basic summary statistics of the attributes

Most attributes exhibit a symmetric distribution around the median, indicating balanced data with no severe skewness. The median is generally well-centered between the min and max, suggesting consistency in the data. However, several attributes, such as malic acid and color intensity show more obvious variability across the data. Where the standard deviation leans high like in the case of malic acid (1.12) and a wide range of 0.74 to 5.80, showing significant diversity among the wine samples.

The range and variability differ across attributes. Proline, color intensity, and magnesium have wider ranges, indicating greater variability among samples, while nonflavonoid phenols and proanthocyanins display narrower ranges, suggesting more uniform data. Despite this, the data is generally consistent, with most values within expected ranges, indicating reliability.

The scatter plots for Malic acid, Magnesium, Ash, and Color intensity show significant outliers, indicating notable variability in these attributes. Malic acid has high outliers above

	Alcohol	Malic acid	Ash	Alcalinity of ash	Magnesium	Total phenols	Flavanoids	Nonflavanoid phenols	Proanthocyanins	Color intensity	Hue	OD280/OD315 of diluted wines	Proline
mean	12.99	2.34	2.37	19.52	99.59	2.29	2.02	0.36	1.59	5.05	0.96	2.60	745.10
std	0.81	1.12	0.28	3.34	14.17	0.63	1.00	0.12	0.57	2.32	0.23	0.71	314.88
min	11.03	0.74	1.36	10.60	70.00	0.98	0.34	0.13	0.41	1.28	0.48	1.27	278.00
25	12.36	1.60	2.21	17.20	88.00	1.74	1.20	0.27	1.25	3.21	0.78	1.93	500.00
50	13.05	1.87	2.36	19.50	98.00	2.35	2.13	0.34	1.55	4.68	0.96	2.78	672.00
75	13.67	3.10	2.56	21.50	107.00	2.80	2.86	0.44	1.95	6.20	1.12	3.17	985.00
max	14.83	5.80	3.23	30.00	162.00	3.88	5.08	0.66	3.58	13.00	1.71	4.00	1680.00
median	13.05	1.87	2.36	19.50	98.00	2.35	2.13	0.34	1.55	4.68	0.96	2.78	672.00

Table 3: Summary statistics of wine dataset

4; Magnesium shows clear outliers above 140, highlighting substantial deviations, possibly from soil differences. Ash displays a few mild outliers above 3.0, with a relatively uniform distribution. Color intensity has pronounced outliers above 10, indicating a distinct group with higher levels.

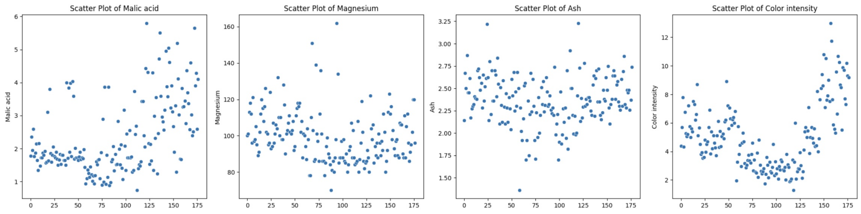

Figure 1: *Attributes Outliers Analysis*

Further investigation is needed to determine if these outliers reflect natural variability or errors. Analyzing correlations between attributes could provide additional insights, as strong correlations might reveal that outliers in one attribute are linked to variations in another.

3 Data visualization and PCA

3.1 Bad values in the data

A test using Pandas library to check for NaN and non-numeric values was run and it was found that the data contains **no missing values**, **no corruptions** or **incorrect data types**.

3.2 Data distribution

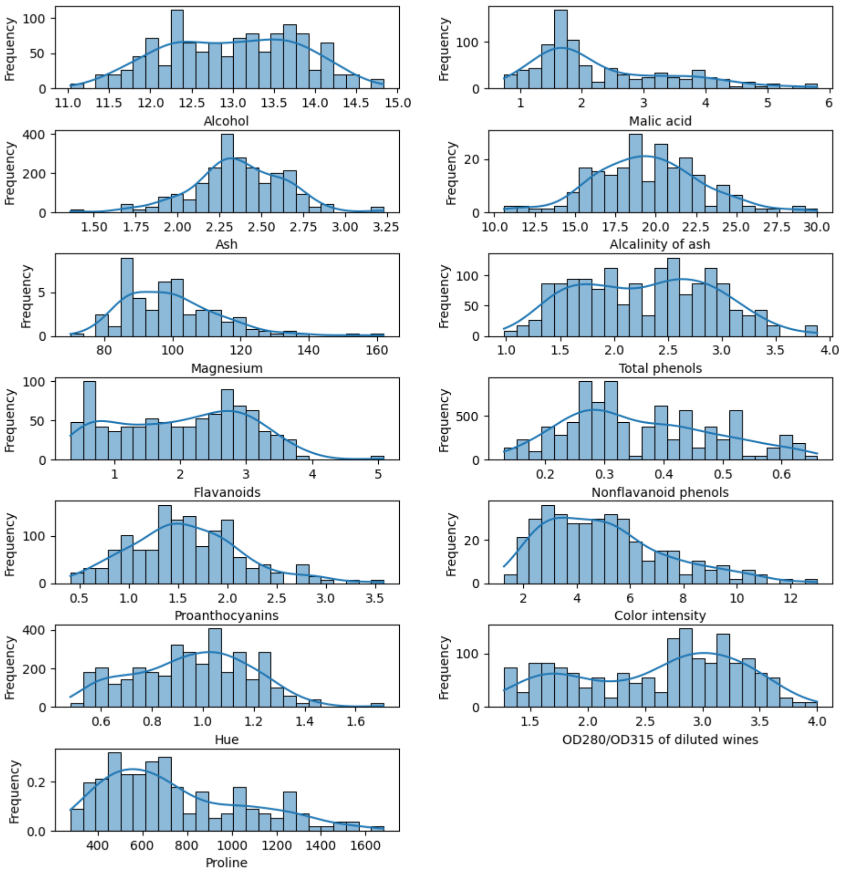

Figure 2: Wine dataset attributes Kernel Density Estimates plotted showing the frequency of occurrence of values

Kernel Density Estimates (Figure 2) allow us to visually show the distribution of the data. Frequency of occurrence within the bins (20 bins) was used as a way to be outlier invariant. We can see that most attributes follow fairly clean normal distributions with some natural skew. Attributes that do not from a visual inspection include Flavanoids, Proline and OD280/OD315 of diluted wines.

To be able to mathematically label and state which attributes follow a normal distribution we also performed a skewness and kurtosis test (Figure 3). The results indicate that six of the attributes are approximately normally distributed:

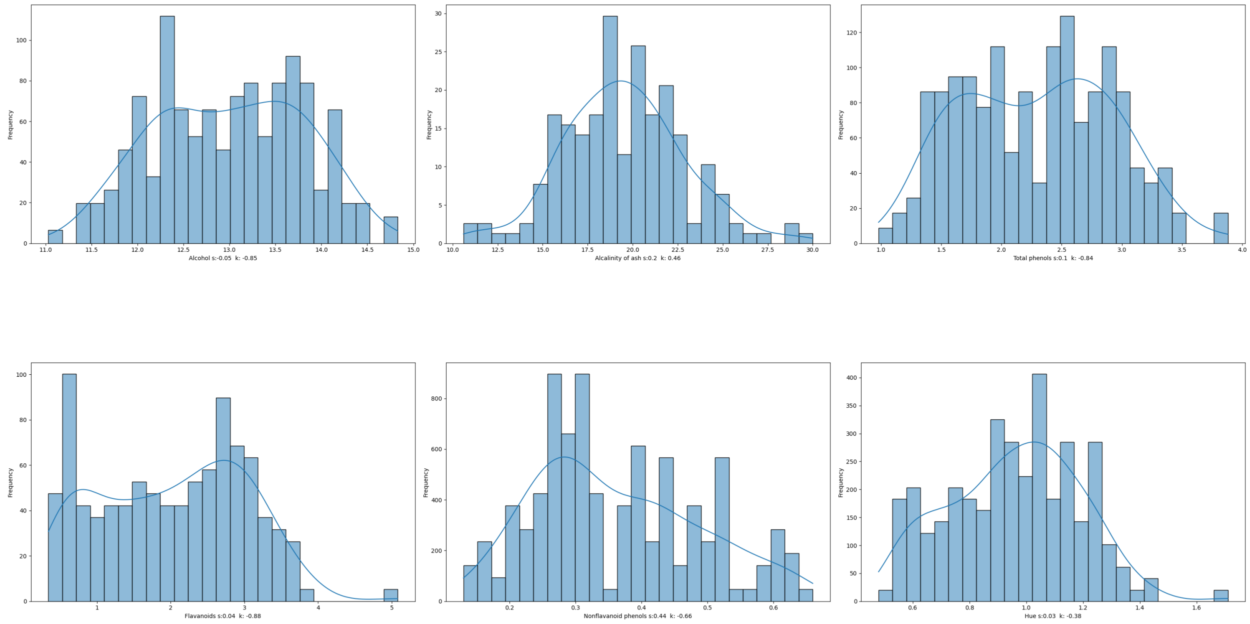

Figure 3: *Attributes distribution*

3.3 Correlations between the different attributes of the Wine dataset.

To clearly and concisely visualise correlation between the data, a Pearson correlation test was run where it looks for linear correlation between two datasets or in this case attributes, using normalised measurement of covariance to ensure that the result always has a value between -1 and 1. Looking at the output we find that the most positively correlated attributes include Flavonoids and Total phenols, Total phenols and OD280/OD315, Alcohol and Proline and finally Flavonoids and OD280/OD315. A positive and significant correlation tells us as one attribute value increases so does the other attribute. Drawing attention to the strongest positive correlation between Flavonoids and Total phenols, the relationship can be simply explained by the fact that Flavonoids are a subset of phenolic compounds [3] which contribute to the wines taste, structure and stability. Another interesting significant correlation is between Proline and Alcohol content. As grapes ripen, proline levels increase, and since ripening is also linked to sugar content (which ferments into alcohol), there's a natural relationship between proline and alcohol concentration in wine [4]. Given these known correlations within the attributes allows us to correct redundancies, leading to more efficient models with less risk of overfitting.

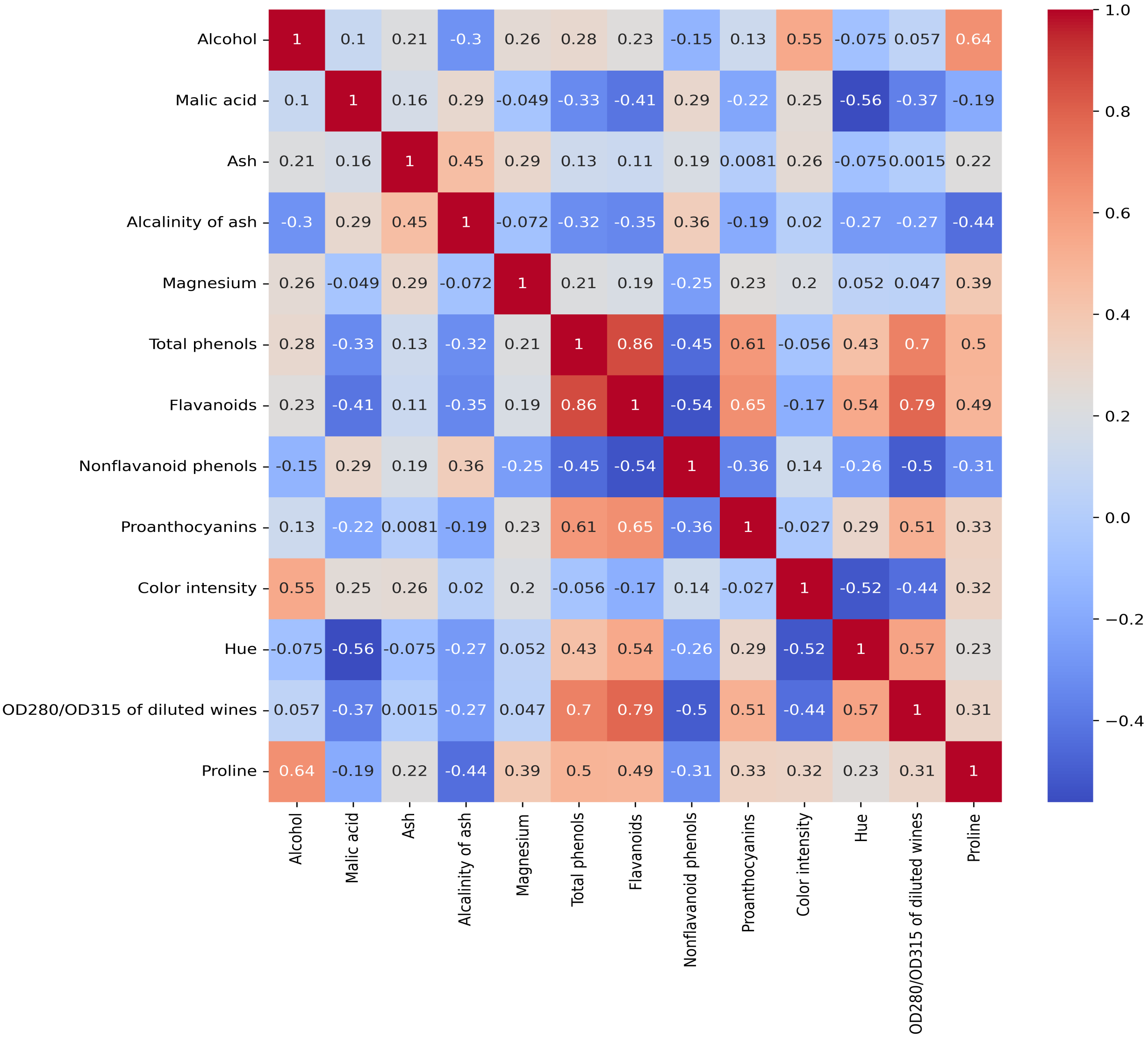

Figure 4: Correlation matrix of all the attributes in the Wine dataset

3.4 The PCA analysis

3.4.1 Variation

The function demonstrates that 7 to 8 principal components (PCA) are needed to explain 90% or more of the data's variance. While it's possible to visualize the data using only 2 or 3 principal components, they account for less than 60% and 80% of the total variance, respectively.

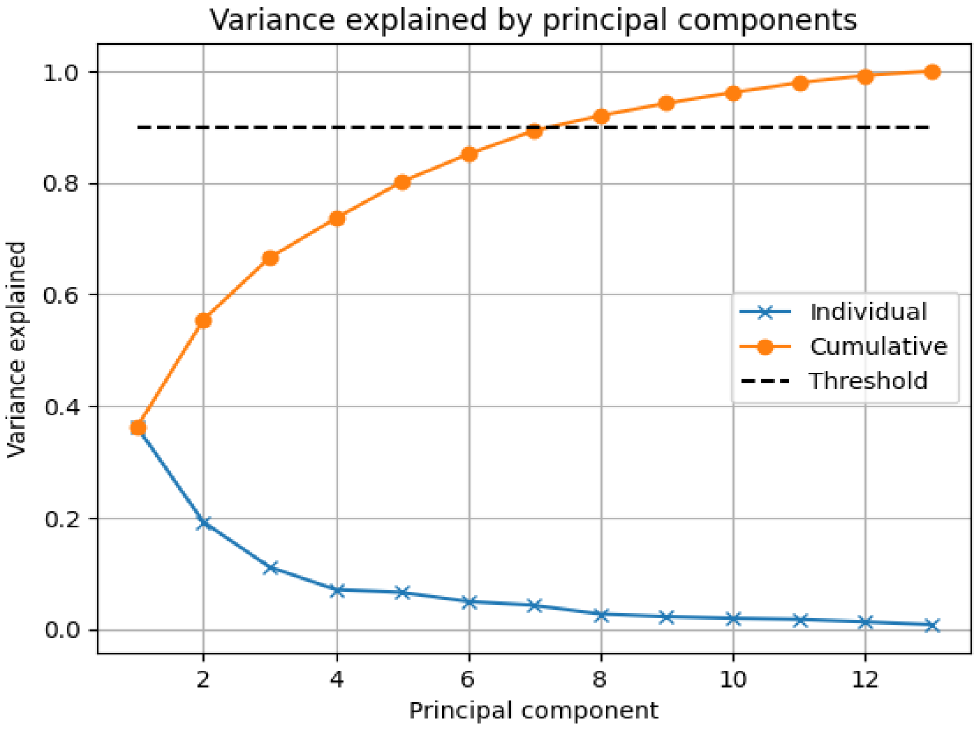

Figure 5: Wine dataset attributes Kernel Density Estimates plotted showing the frequency of occurrence of values

3.4.2 PCA components

To better visualize the data in two dimensions, PCA1 and PCA2 are used. However, these two principal components explain only 60% of the variance. By extending the visualization to three dimensions and including a third principal component (PCA3), the explained variance increases by 20%, potentially providing more insight into the data.

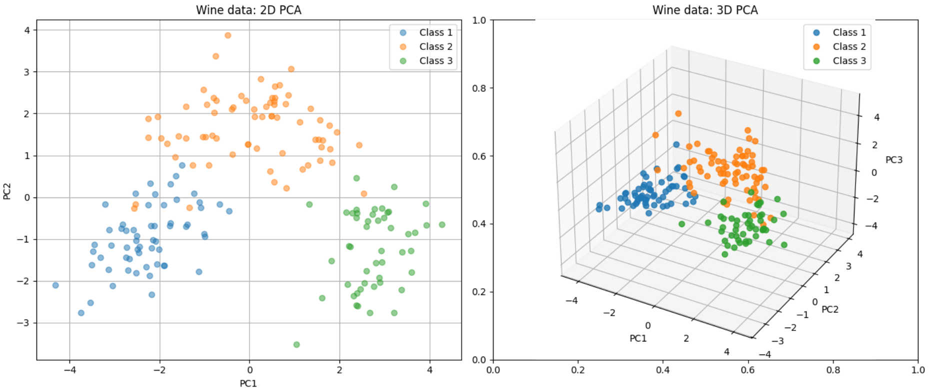

Figure 6: data plotted with 2 and 3 PCA

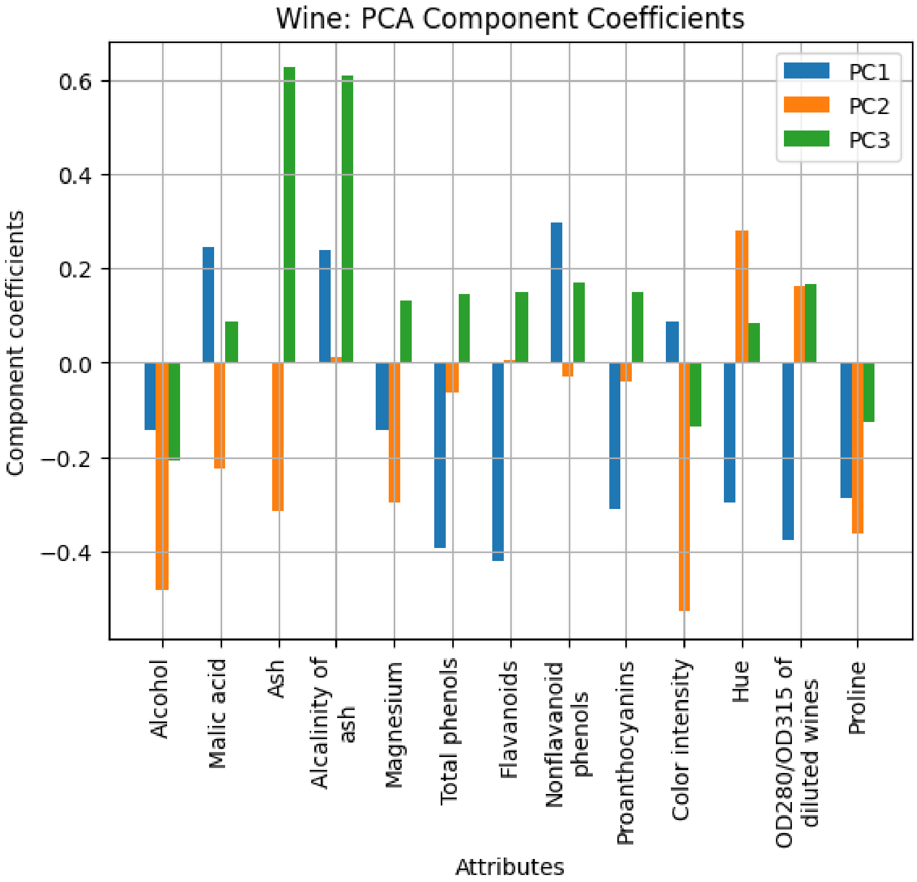

Figure 7: Attributes and plot counterparts

In the context of the first PCA for wine characterization, the value derived from various chemical attributes reflects the overall quality and characteristics of the wine. The value decreases (becomes more negative) when the wine has higher levels of the following attributes: Alcohol, Magnesium, Total Phenols, Flavonoids, Proanthocyanins, Hue, OD280/OD315 of Diluted Wines and Proline. An increase in these components suggests a wine that is potentially richer, more robust, or more complex, contributing to a lower overall value in this specific PCA component framework.

Conversely, the value increases when the wine has higher levels of the following attributes: Malic Acid, Nonflavonoid Phenols, Color Intensity and Alkalinity of Ash.
Higher concentrations of these components generally indicate a wine that is fresher, more vibrant, or lighter in style, leading to a higher value in this PCA component context.

The second PCA component reflects the overall quality and characteristics of the wine based on various chemical attributes. The second PCA component increases with higher levels of the following attributes: Alcalinity of Ash - moderate increase, Hue (A10) - relevant increase and OD280/OD315 of Diluted Wines (A11) - relevant increase. These attributes generally suggest a wine that exhibits certain desirable characteristics, contributing positively to the second PCA component. Conversely, the second PCA component decreases (becomes more negative) with higher levels of the following attributes: Alcohol, Color Intensity, Proline, Total Phenols and Flavonoids. An increase in these components suggests a wine that may be heavier or less vibrant, resulting in a lower value for the second PCA component.

In addition to the second PCA component, the third PCA component also reflects the characteristics of the wine based on various chemical attributes. It increases with higher levels of Malic Acid, Ash, Alcalinity of Ash, Magnesium, Total Phenols, Flavonoids, Non-flavonoid Phenols, Proanthocyanins, Hue and OD280/OD315 of Diluted Wines. Conversely, the third PCA component decreases (becomes more negative) with higher levels of the following attributes: Alcohol, Color Intensity and Proline.

3.5 Key lessons learned

Based on the visualization performed, the primary machine learning modeling aim of classification appears to be feasible. The visualizations indicate that the data is well-distributed and consistent, with manageable outliers and significant attribute variability, which are crucial for effective classification. Attributes such as proline, color intensity, and magnesium show wide ranges, highlighting the diversity in the dataset that aids in distinguishing between classes.

We normalized the data by removing the mean and dividing the standard deviation, prior to applying PCA because the attributes in the dataset are measured on different scales. For instance, the values of proline and magnesium differ significantly in magnitude compared to other attributes. Without normalization, attributes with larger numerical ranges, like proline, would dominate the PCA results, leading to principal components that are not meaningful. Normalizing ensures that all attributes contribute equally to the analysis, allowing PCA to accurately capture the underlying structure of the data.

Furthermore, correlations between attributes like Flavonoids and Total phenols provide useful insights for predictive modeling. Also, the PCA analysis suggests that essential data characteristics can be efficiently captured in reduced dimensions, facilitating the classification process. Overall, the visual evidence supports the successful application of machine learning models to reliably categorize the data both using regression for components like alcohol percentage and classification for the wine cultivar.

4 Problems Resolution (Contribution: Equal split)

1. **Problem 1.**

   (a) Answer = D.

   (b) This is because Time of Day is NOT nominal, ratio or ordinal allowing us to remove those potential answers.

2. **Problem 2.**

   (a) Answer = A.

   (b) This is because given vectors:

$$\mathbf{x}_{14} = \begin{bmatrix} 26 \\ 0 \\ 2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_{18} = \begin{bmatrix} 19 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

   The ∞-norm (maximum norm) distance is calculated as:

$$d_\infty(\mathbf{x}_{14}, \mathbf{x}_{18}) = \max(|26 - 19|, |0 - 0|, |2 - 0|, |0 - 0|, |0 - 0|, |0 - 0|).$$

$$d_\infty = \max(7, 0, 2, 0, 0, 0) = 7.$$

3. **Problem 3.**

   (a) Answer = A

   (b) The total variance is the sum of the squares of the singular values:

$$\text{Total Variance} = 13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2 = 670.37.$$

   Variance explained by the first four principal components:

$$\text{Variance}_{1+2+3+4} = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{670.37} = \frac{581.07}{670.37} \approx 0.866.$$

4. **Problem 4.**

   (a) Answer = D, is the answer because our three high values lie on positive components of PCA2.

   (b) A, cannot be the answer because a high value in the attr 3 and 4 will lead to a negative projection. B, can also not be the answer as a high value in attr 4 would lead to a large negative projection. C, is also not the answer as the high value lies on the positive component of the attr which would require a positive project.

5 References

[1] Comparative analysis of statistical pattern recognition methods in high dimensional settings. Stefan Aeberhard, Danny Coomans, Olivier de Velhttps://doi.org/10.1016/0031-3203(94)90145-7

[2] Li, Y., Tang, Z. and Yao, J. (2023). DEPSOSVM: An Alternative Wine Classification Method Based on Machine Learning. Inteligencia Artificial, 26(71), 131–141. https://doi.org/10.4114/

[3] Zoecklein, B.W., Fugelsang, K.C., Gump, B.H., Nury, F.S. (1995). Phenolic Compounds and Wine Color. In: Wine Analysis and Production. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-6978-4_7

[4] Espinase Nandorfy, D., Likos, D., Lewin, S., Barter, S., Kassara, S., Wang, S., Kulcsar, A., Williamson, P., Bindon, K., Bekker, M., Gledhill, J., Siebert, T., Shellie, R.A., Keast, R., Francis, L., Krstic, M. (2024). Blending benefits from high-proline wines. Wine and Viticulture Journal, 39(1), 33-35.