Cherkashina Elizaveta

Machine Learning Final Project

Report

Introduction

Understanding the chemical composition of wine has profound implications for both quality control and the study of viticulture. In this project, we investigate whether the class of a wine—a botanical grouping corresponding to three different grape cultivars—can be reliably predicted from 13 routinely measured chemical attributes. We further examine whether unsupervised clustering uncovers the same natural divisions among samples without any prior labeling.

Data and Exploratory Analysis

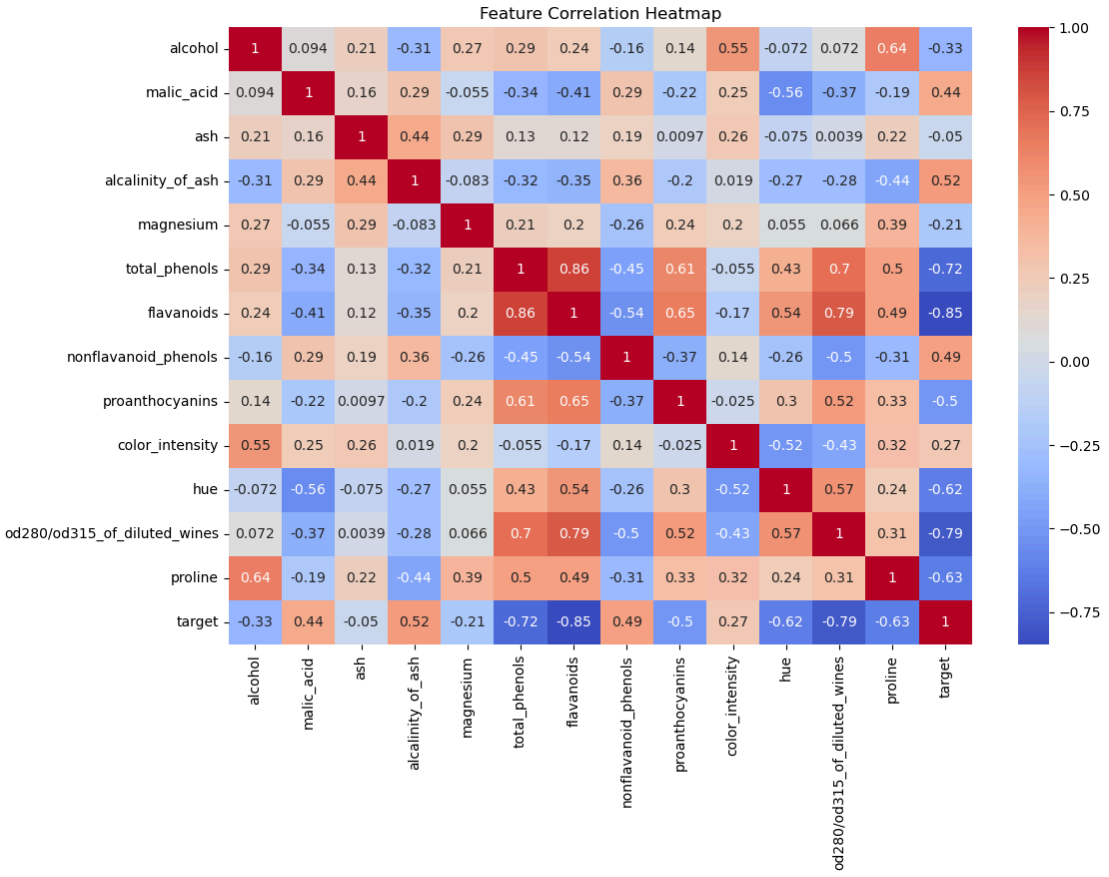
The dataset originates from scikit-learn and comprises 178 observations. Each sample is described by 13 continuous variables such as alcohol content, flavonoid concentration, and spectrophotometric readings (e.g. OD280/OD315). The first step involved generating summary statistics, which revealed a wide range of scales across features (for example, Proline ranged from approximately 300 to 1,700, while flavonoids varied between 0.7 and 5).

Next, we computed a Pearson correlation heatmap of all features against each other and against the target class. Two notable patterns emerged:

- Strong collinearity between total phenols and flavonoids (correlation ≈ 0.86), which can be problematic;

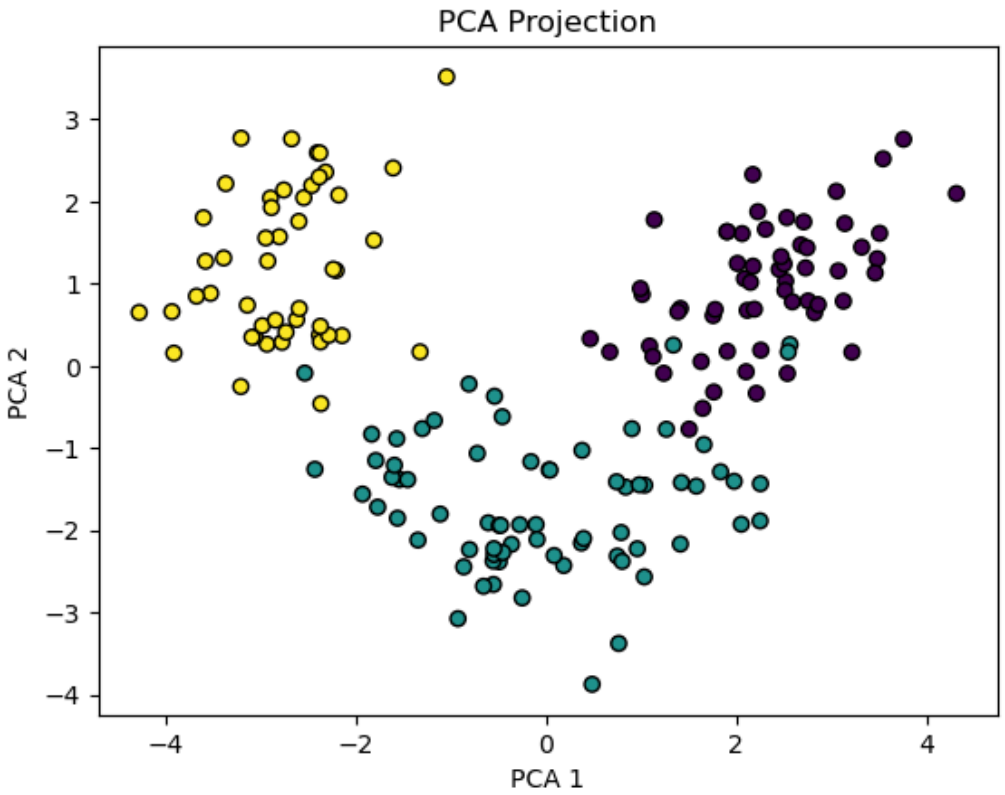
- Moderate correlation of attributes like flavonoids, OD280/OD315, and color intensity with the wine class.

These insights guided our preprocessing: all features were standardized (zero mean, unit variance) to ensure comparability in the modeling phase.



Dimensionality Reduction with PCA

To visualize the complex 13-dimensional space, we applied Principal Component Analysis (PCA). The first two principal components captured approximately 55% of the total variance. Plotting the first component against the second revealed that samples from class 0 and class 2 formed largely distinct clouds, while class 1 overlapped partially with both groups. This preliminary result suggested that linear methods might suffice for classification and offered a clear visual separation among most classes.



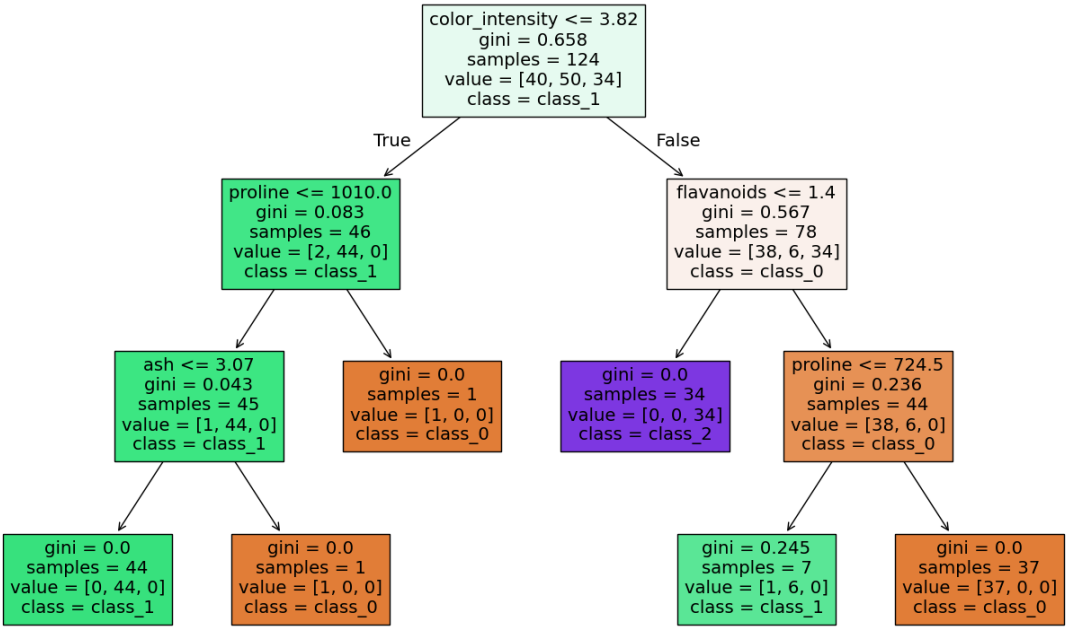
Supervised Learning

1. **Train/Test Split**

   We split the data into 70% training and 30% testing sets, using a fixed random seed for reproducibility. Before settling on a 70/30 split, 40/60 and 20/80 splits were also tested, but did not produce better results than the chosen split.

2. **Decision Tree Classifier**

   The Decision Tree model has 96% accuracy, misclassifying one instance of class 0 and one instance of class 2 as class 1. Considering that class 1 was overlapping with other classes when plotting, it can be difficult for the model to distinguish instances that are overlapping. The maximum depth of the tree set to 3 was also determined through iterations; depths above or below were producing more misclassifications.

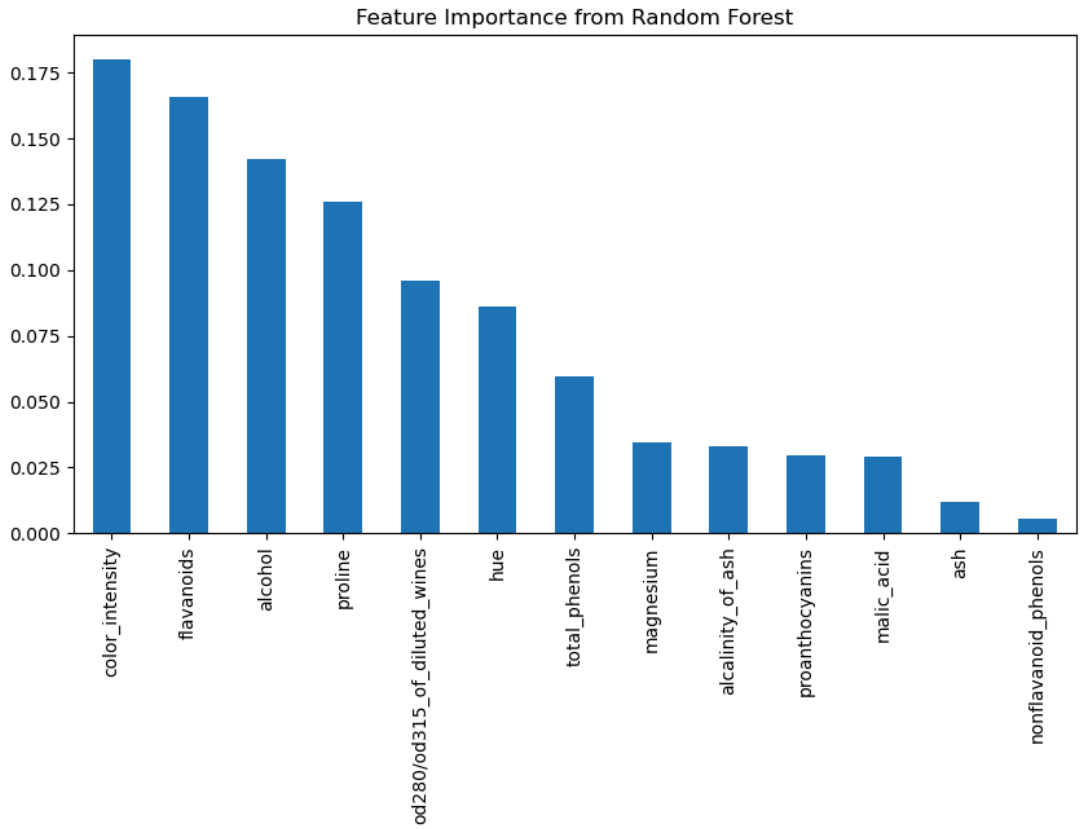


	precision	recall	f1-score	support
class_0	1.00	0.95	0.97	19
class_1	0.91	1.00	0.95	21
class_2	1.00	0.93	0.96	14
accuracy			0.96	54
macro avg	0.97	0.96	0.96	54
weighted avg	0.97	0.96	0.96	54

3. **Random Forest Classifier**

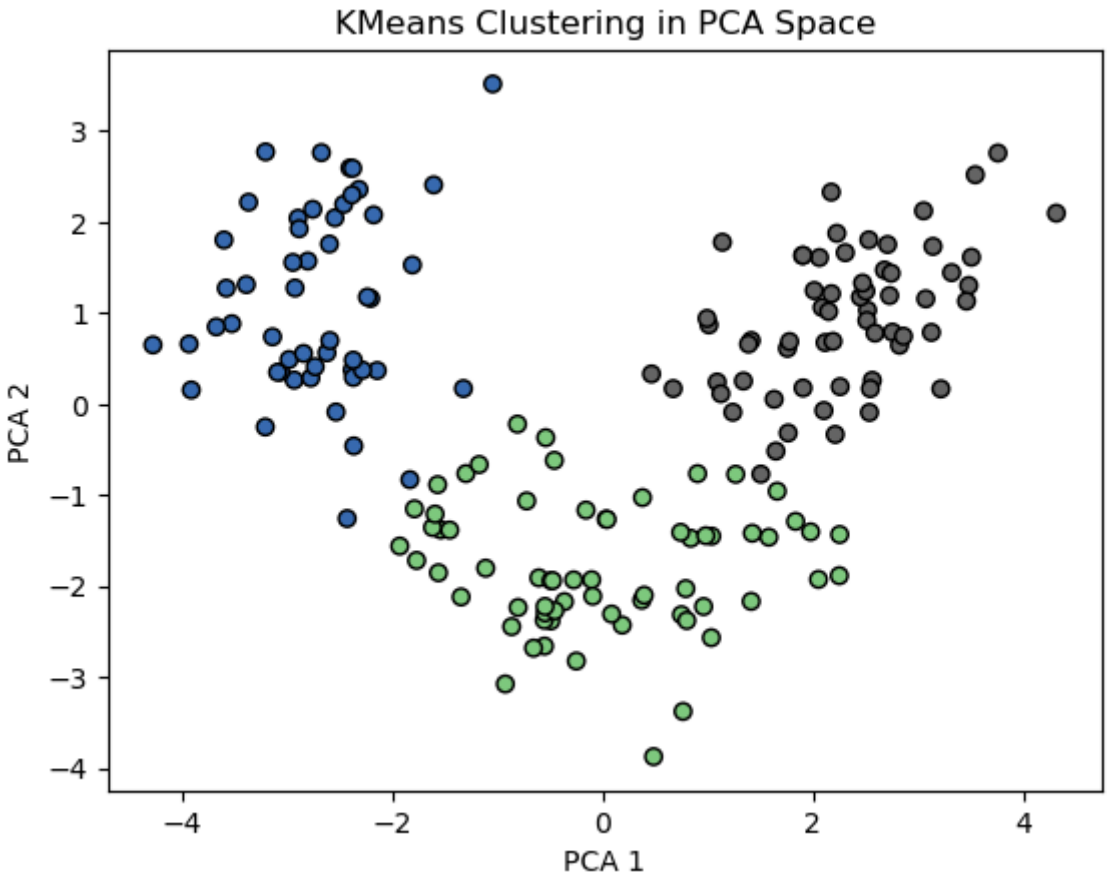
   The Random Forest model achieved 100% accuracy on the test set. This can be explained by the reduction of overfitting and bagging that characterize Random Forest models and help produce better generalization in comparison to Decision Trees. Feature-importance scores ranked color intensity, flavonoids, and alcohol as the top predictors, confirming the key role of phenolic compounds in distinguishing cultivars.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	19
1	1.00	1.00	1.00	21
2	1.00	1.00	1.00	14
accuracy			1.00	54
macro avg	1.00	1.00	1.00	54
weighted avg	1.00	1.00	1.00	54



Unsupervised Learning

To test whether the natural structure of the data aligns with the known labels, we applied K-Means clustering (k = 3) on the scaled features. When projected into the same two PCA dimensions, the resulting clusters overlapped strongly with the true classes. Computing the Adjusted Rand Index between cluster labels and true labels yielded approximately 0.89, indicating high agreement and confirming that the chemical measurements alone naturally partition the samples into their botanical origins.



Discussion and Conclusion

The analysis demonstrates that simple chemical assays can robustly distinguish among grape cultivars. The Random Forest's perfect accuracy and its feature‑importance rankings highlight that phenolic compounds (flavonoids, total phenols), as well as color characteristics, are the most discriminative. Unsupervised clustering corroborates these findings by revealing that the same attributes drive the natural grouping of samples even without label information.

Conclusion

The answer to the research question is yes—the botanical origin of wine can be predicted with high accuracy from its chemical profile and other features, and the inherent chemical differences among cultivars form clear clusters even in an unsupervised setting.

Reflections and Future Work:

While this dataset is relatively small and sourced from a single region, the results suggest broader applications in quality control and varietal authentication. Future studies

might expand to more diverse vintages, explore nonlinear embedding techniques (such as t-SNE or UMAP), or validate models on external datasets to assess generalizability.