DSCI 310: Predicting Wine Cultivars

Zhibek Dzhunusova, Andrea Jackman, Kaylan Wallace & Chuxuan Zhou

Table of contents

Summary

In this project, we will predict what cultivar a wine was derived from based on its chemical properties.

The data was sourced from the UCI Machine Learning Repository (Aeberhard and Forina 1991). It contains data about various wines from Italy derived from three different cultivars. Each row represents the chemical and physical properties of a different wine, such as its concentration of alcohol, magnesium level and hue.

Introduction

Wine is a beverage that has been enjoyed by humans for thousands of years (Fehér, Lengyel, and Lugasi 2007). Consequently, humans have a long agricultural history with the grape plant which has led to the development of many different cultivars: grape plants selected and breed for their desirable characteristics (Harutyunyan and Malfeito-Ferreira 2022). Our dataset contains information about twelve chemical properties of 178 red wines made from three grape cultivars in Italy.

The recorded chemical properties include:

1. Alcohol content

2. Malic acid (gives the wine a fruity flavour)
3. Ash (left over inorganic matter from the wine-making process)
4. Alkalinity of ash (ability to resist acidification)
5. Magnesium, total phenols (contribute to bitter flavour of wine)
6. Flavanoids (antioxidants that contribute to bitter flavour and aroma of wine)
7. Nonflavanoid phenols (weakly acidic)
8. Proanthocyanins (bitter smell)
9. Color intensity
10. Hue
11. The ratio of OD280 to OD315 of diluted wines (protein concentration)
12. Proline (main amino acid in wine, important aspect of the flavour) (Bai, Wang, and Li 2019).

Using this dataset, our predictive question is: "What is the cultivar of an unknown wine based on the chemical properties?"

Identifying the chemical properties that distinguish cultivars enables farmers to make informed decisions about grape cultivation, aligning grape varieties with desired wine characteristics. By selecting cultivars known for specific flavor profiles or chemical compositions, farmers can tailor vineyard practices to meet market demands effectively. Moreover, this knowledge empowers brewers to experiment with wine compositions, fostering innovation and the creation of novel flavors. Armed with a deep understanding of wine chemistry, brewers can also strategically market their products, ensuring effective communication of the unique qualities and appeal of each wine to consumers.

Exploratory Data Analysis

In Table 1, we have summarized the mean, maximum, minimum and standard deviation for all predictors. This gives us a better idea of the normal range of values for each predictor within our model.

Figure 1 depicts the distribution of proline and ash values for each cultivar. Cultivar 1 has more distinct proline and ash values while Cultivar 2 and Cultivar 3 overlap more substantially.

Figure 2 shows distribution of alcohol content for the wines from each cultivar. We can see that each cultivar has a narrow range of values that wines tend to fall within which is relatively distinct for each cultivar. This means this could be an effective predictor of cultivar.

Methods & Results

This project utilized a K-nearest neighbours algorithm to predict what cultivar a wine was derived from based on its various chemical properties. First, we read in data from the UCI Machine Learning Repository. It contains data about various wines from Italy derived from

Table 1: Summary statistics for the raw data.

alcohol_mean	alcohol_sd	alcohol_min	alcohol_max
13.00062	0.8118265	11.03	14.83

malicacid_mean	malicacid_sd	malicacid_min	malicacid_max
2.336348	1.117146	0.74	5.8

ash_mean	ash_sd	ash_min	ash_max
2.366517	0.274344	1.36	3.23

alcalinity_of_ash_mean	alcalinity_of_ash_sd	alcalinity_of_ash_min	alcalinity_of_ash_max
19.49494	3.339564	10.6	30

magnesium_mean	magnesium_sd	magnesium_min	magnesium_max
99.74157	14.28248	70	162

total_phenols_mean	total_phenols_sd	total_phenols_min	total_phenols_max
2.295112	0.625851	0.98	3.88

flavanoids_mean	flavanoids_sd	flavanoids_min	flavanoids_max
2.02927	0.9988587	0.34	5.08

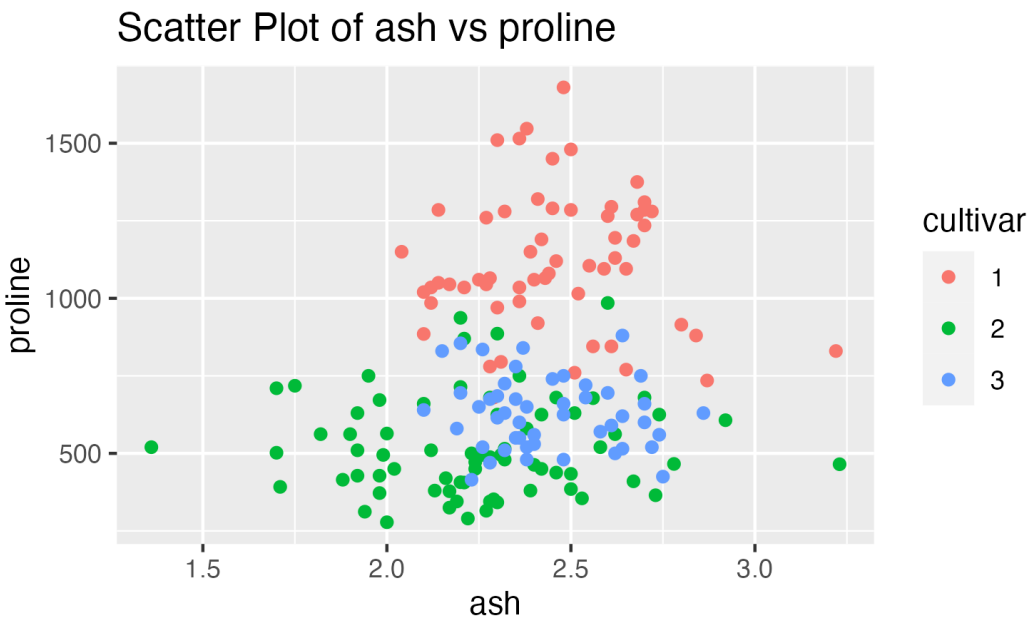

Figure 1: Scatterplot of proline and ash values for each cultivar

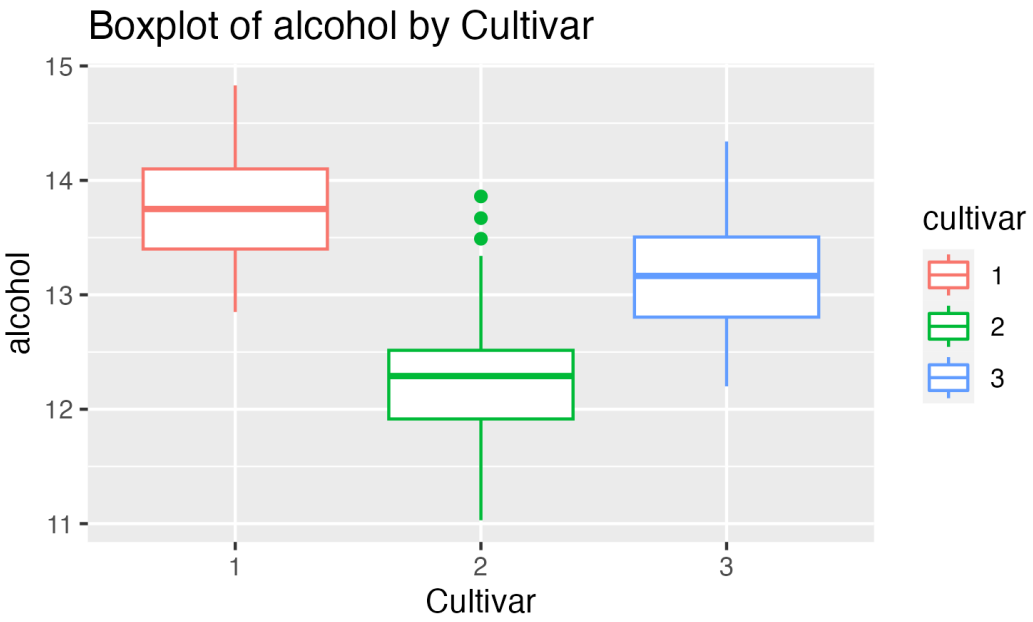

Figure 2: Boxplotplot of alcohol content for each cultivar

three different cultivars. Each row represents the chemical and physical properties of a different wine, such as its concentration of alcohol, magnesium level and hue.

We then tidied the data and balanced the classes of the classification variable we are interested in. This is because the data set is not extensively large, so ensuring each class has an equal number of observations prevents our model from being biased towards a specific dominant class. Next we calculated some summary statistics to facilitate exploratory data analysis, with the goal of finding key input variables for our model.

The data was split into 75 for the training set and 25 for the test set. To fine tune our model, we used 5 fold cross validation, grid search, and graphical methods to choose the optimal value of K. The result was $K = 8$ being used in the k-nn model.

Discussion

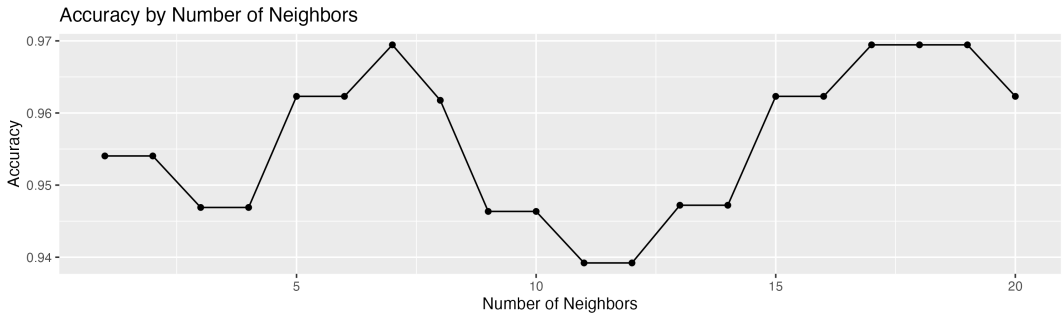

Figure 3: Model accuracy as a function of number of neighbors

We can see from Figure 3 that accuracy remains high across from k = 1 to k = 20. It peaks at around 0.98 for $K = 8$ before decreasing and subsequently increasing back to 0.98 from $K = 17$ to $K = 19$. This is a high accuracy value and will increase the power of our predictive model.

Table 2: Model Evaluation metrics.

Prediction	Truth	n
1	1	15
2	1	0
3	1	0
1	2	1
2	2	17
3	2	0
1	3	0
2	3	0

Table 2: Model Evaluation metrics.

Prediction	Truth	n
3	3	12

Table 2 shows how well our model is able to predict the cultivar type from predictor variables. We see that it accurately predicts cultivar for 44 of 45 data points and only mistakes cultivar 1 for cultivar 2 once. Therefore, our model has a very high success rate and will be able to accurately predict the cultivar in most cases.

Our multiclass k-nn model performed relatively well on the test data, achieving an accuracy estimate of approximately 0.98. The confusion matrix reveals insights into the model's performance across the three cultivar classes. Notably, while the model demonstrated strong precision and recall for predicting cultivar 3, it encountered challenges in accurately classifying cultivar 2. This aligns with our initial hypothesis that certain chemical properties may serve as distinguishing factors for wine cultivars.

However, despite the model's overall success, its limitations in predicting cultivar 2 suggest avenues for improvement. Future iterations of the model could benefit from refining input variables to better capture the nuances of each cultivar's chemical composition. Moreover, our findings underscore the importance of further investigation into the unique characteristics of cultivar 3, which consistently stood out in our predictions.

By elucidating the chemical properties that differentiate wine cultivars, our study contributes to the broader goal of simplifying wine classification for consumers. Ultimately, this research not only enhances our understanding of wine chemistry but also has practical implications for wine enthusiasts and industry professionals alike.

References

Aeberhard, Stefan, and M. Forina. 1991. "Wine." https://doi.org/10.24432/C5PC7J. https://doi.org/10.24432/C5PC7J.

Bai, X., L. Wang, and H. Li. 2019. "Identification of Red Wine Categories Based on Physicochemical Properties." In *International Conference on Educational Technology, Management, and Humanities Science*, 1443–48. https://doi.org/10.25236/etmhs.2019.309.

Fehér, J., G. Lengyel, and A. Lugasi. 2007. "The Cultural History of Wine—Theoretical Background to Wine Therapy." *Central European Journal of Medicine* 2 (4): 379–91. https://doi.org/10.2478/s11536-007-0048-9.

Harutyunyan, M., and M. Malfeito-Ferreira. 2022. "The Rise of Wine Among Ancient Civilizations Across the Mediterranean Basin." *Heritage* 5 (2): Article 2. https://doi.org/10.3390/heritage5020043.