DSCI 310: Predicting Wine Cultivars

Zhibek Dzhunusova, Andrea Jackman, Kaylan Wallace & Chuxuan Zhou

Table of contents

Summary

In this project, we asked whether we can 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 178 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.

Using a k-nearest neighbors algorithm we attempted to predict the cultivar of an unknown wine based on physiochemical properties such as alcohol content and magnesium levels. Our model achieved 98% accuracy on our test data, but struggled to accurately predict the origin of wines from cultivar 2. This demonstrates that most cultivars are strongly distinguished by their physiochemical properties with some overlap between similar cultivars. Our analysis effectively identified chemical properties that strongly distinguish cultivars. This provides valueable information for winemakers about which grapes to cultivate to achieve their desired wine characteristics and informs future innovation in the wine industry.

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 (Aeberhard and Forina 1991).

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.

Table 1: Summary statistics for the raw data.

	alcohol	malicacid	ash	alcalinity_of_ash
min	11.0300000	0.740000	1.360000	10.600000
max	14.8300000	5.800000	3.230000	30.000000
mean	13.0006180	2.336348	2.366517	19.494944
sd	0.8118265	1.117146	0.274344	3.339564

	magnesium	total_phenols	flavanoids	nonflavanoid_phenols
min	70.00000	0.980000	0.3400000	0.1300000
max	162.00000	3.880000	5.0800000	0.6600000
mean	99.74157	2.295112	2.0292697	0.3618539
sd	14.28248	0.625851	0.9988587	0.1244533

	proanthocyanins	color_intensity	hue	X0D280_0D315_ratio
min	0.4100000	1.280000	0.4800000	1.2700000
max	3.5800000	13.000000	1.7100000	4.0000000
mean	1.5908989	5.058090	0.9574494	2.6116854
sd	0.5723589	2.318286	0.2285716	0.7099904

	proline
min	278.0000
max	1680.0000
mean	746.8933
sd	314.9075

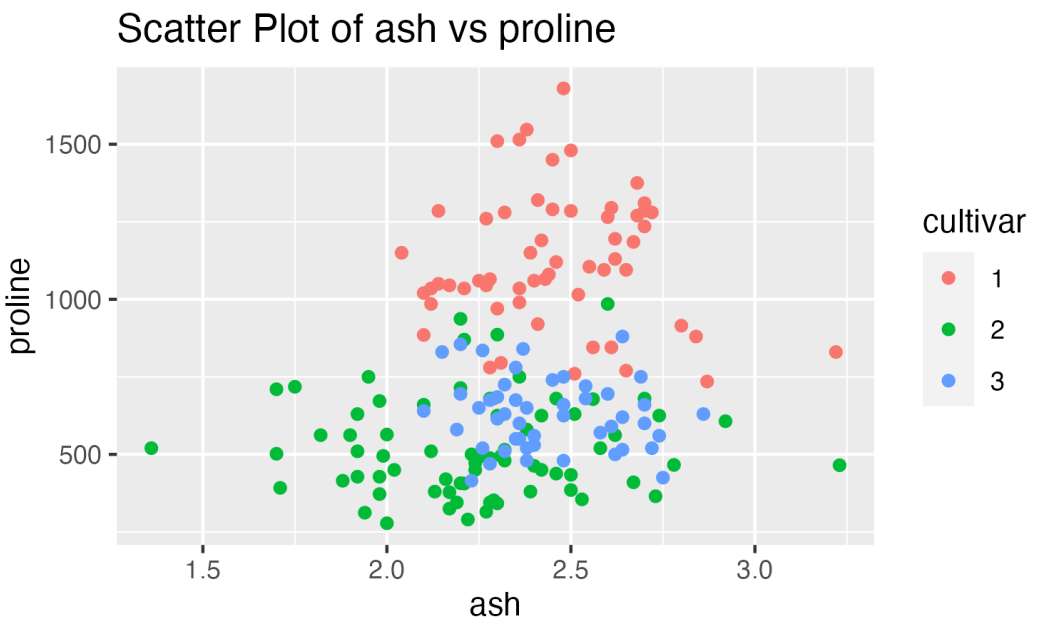

Figure 1: Scatterplot of proline and ash values for each cultivar

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

Data

This project utilized a K-nearest neighbours classification 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 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.

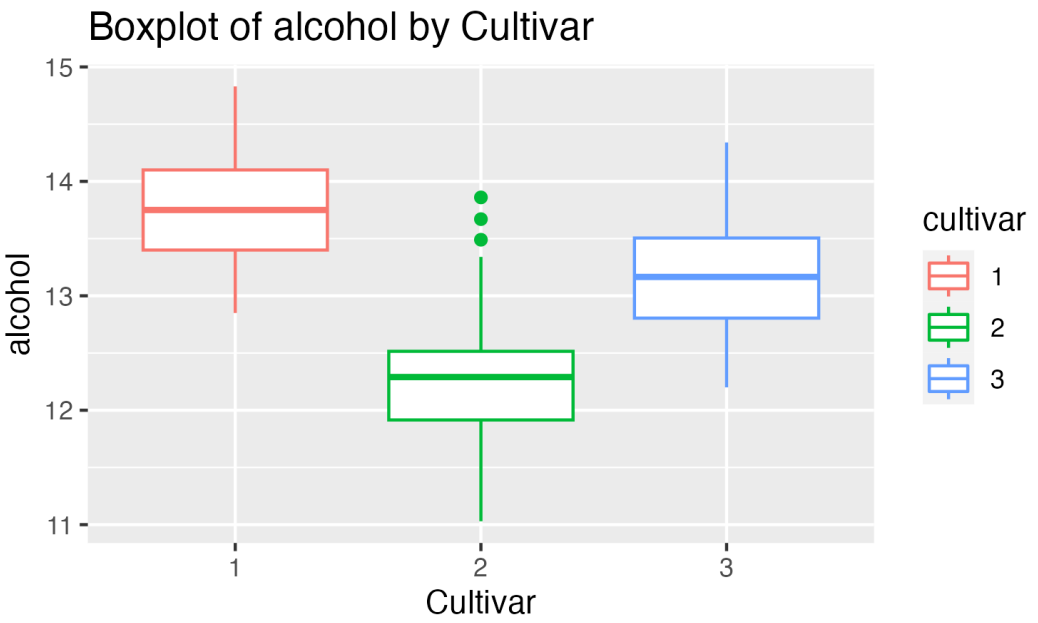

Figure 2: Boxplotplot of alcohol content for each cultivar

Analysis

This project utilized the `R` language to create a K-nearest neighbours classification algorithm for predicting what cultivar a wine was derived from based on its various chemical properties. This dataset includes a target variable named `cultivar`, which is converted into a factor to facilitate classification. After conducting EDA, we elected to include all variables from the original data set to fit the model. The data was split into 75 for the training set and 25 for the test set. The model utilizes a grid search combined with cross-validation to select the optimal value of k. This involves defining a grid of k values to evaluate, and iteratively training and evaluating the model for each value. The selected k value maximizes the model's accuracy on the validation set. We also generated an accuracy plot to visualize the relationship between the number of neighbors (k) and the model's accuracy. The plot serves as a visual aid for selecting the optimal value of k, which corresponds to the point where the model achieves the highest accuracy on the validation data. After choosing k, all variables were centered and scaled before being passed to the model. To evaluate the model's performance, we considered its accuracy and a confusion matrix of the classifications. For futher detail on the model, refer to the analysis script.

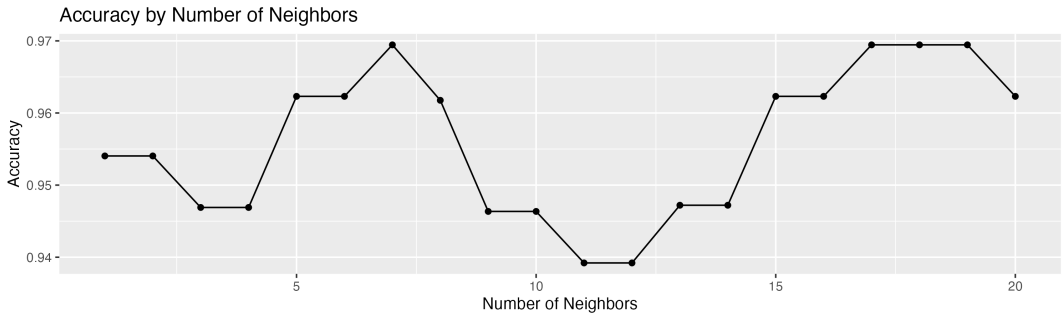

Figure 3: Model accuracy as a function of number of neighbors

Results & Discussion

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
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.