A World of Wine: Exploring Wines Through the Lens of Data

Matthew Bailey
Garrett Kidd
Brenna Wallace
Shaunghnessy Robertson

Introduction

This project aims to use machine learning techniques to build a predictive classification model that can accurately classify wines into two quality categories while also classifying whether the wine is Red or White type. In this presentation, we will explore the different types of wine, the regions where they are produced, and the many factors that can influence their taste and quality.

Inspiration

We, as wine enthusiasts, chose to embark on this project on wine quality because wine has a unique ability to bring people together and create unforgettable experiences. By studying the various factors that contribute to wine quality, we hope to deepen our understanding of this fascinating subject and share our knowledge with others. Through our project, we aim to inspire a greater appreciation for the craftsmanship and dedication of producing high-quality wine and encourage others to explore the world of wine for themselves.

Data

We used two datasets from Kaggle:

- Wine Rating & Price (Red & White Wine)
- Wine Quality Data Set

	Name	Country	Region	Winery	Rating	NumberOfRatings	Price	Year	Type
0	Vermentino 2017	Italy	Toscana	Famiglia Castellani	3.8	25	5.65	2017	white
1	Ronco Broilo 2010	Italy	Colli Orientali del Friuli	Conte d'Attimis Maniago	4.3	25	44.90	2010	white
2	Weisser Schiefer s 2017	Austria	Südburgenland	Weinbau Uwe Schiefer	4.2	25	33.25	2017	white
3	Chardonnay 2018	Germany	Rheinhessen	Krämer - Straight	3.9	25	8.99	2018	white
4	Maganza Zibibbo 2018	Italy	Terre Siciliane	Luna Gaia	3.9	25	8.60	2018	white
...	...	...	...	...	...	...	...	...	...
12425	6th Sense Syrah 2016	United States	Lodi	Michael David Winery	3.8	994	16.47	2016	red
12426	Botrosecco Maremma Toscana 2016	Italy	Maremma Toscana	Le Mortelle	4.0	995	20.09	2016	red
12427	Haut-Médoc 2010	France	Haut-Médoc	Château Cambon La Pelouse	3.7	996	23.95	2010	red
12428	Shiraz 2019	Australia	South Eastern Australia	Yellow Tail	3.5	998	6.21	2019	red
12429	Portillo Cabernet Sauvignon 2016	Argentina	Tunuyán	Salentein	3.4	999	7.88	2016	red

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.45	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.49	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.44	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
4	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.40	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

6497 rows × 13 columns

Data Cleaning

Initially, the data we obtained from both sources was free of any inconsistencies or inaccuracies. The only data manipulation we performed was merging the red and white wine CSV files into a single file to create our Tableau visualizations. During the preprocessing stage, we excluded the "type" column and transformed the "quality" column to reflect binary quality by assigning the values "true" or "false".

Research Questions

Question 1

Can a machine learning model accurately classify wine type using just chemical properties?

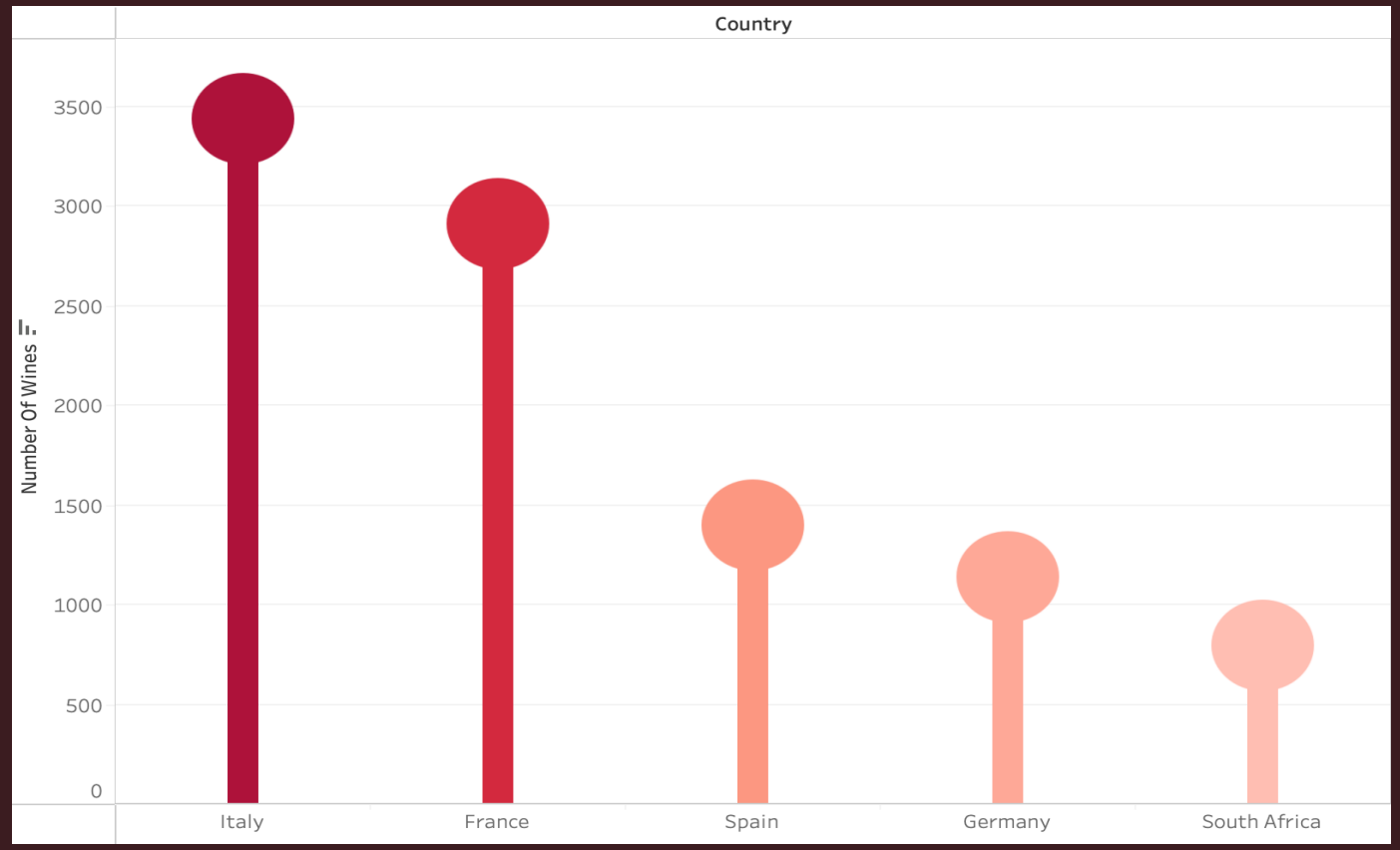
Question 2

Can chemical properties be used to predict if a wine is “good” or “bad” quality?

Question 3

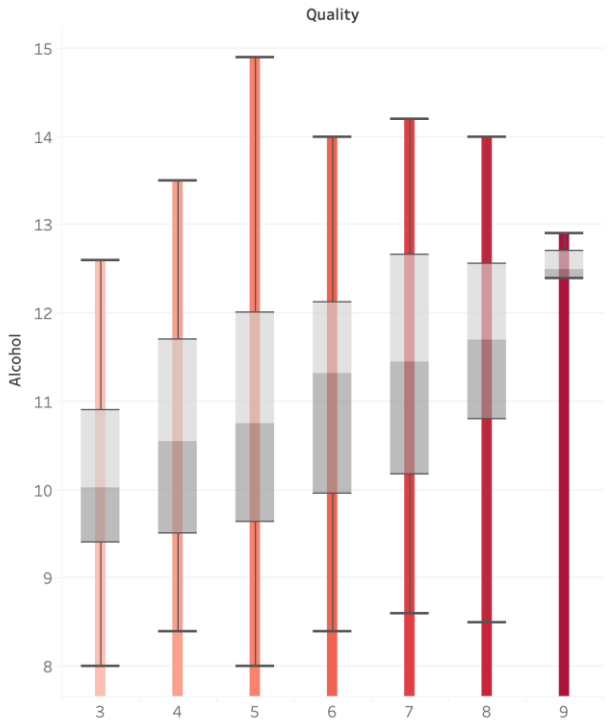
Where are wines produced? Are there countries that produce higher rated or higher priced wine?

Top Wine Producing Countries

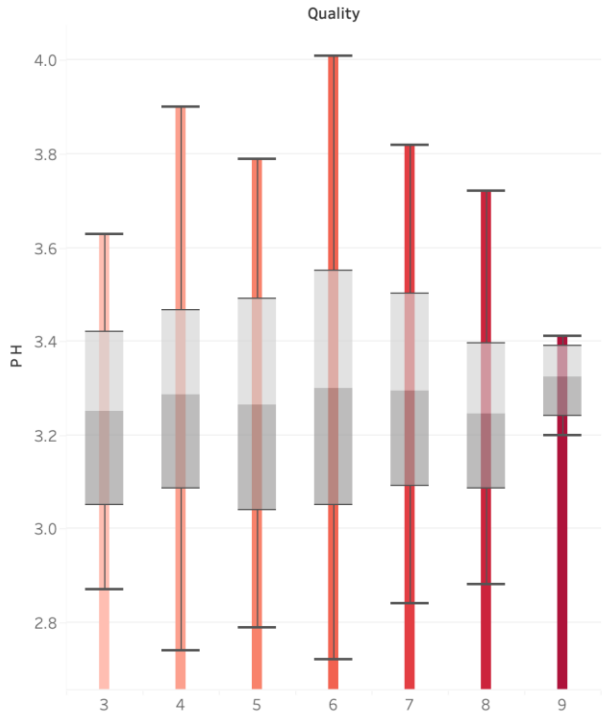


Italy is one of the most diverse and prolific wine-producing countries in the world, with over 3000 different wines. The Italian wine industry is known for its regional diversity, with each region producing unique wines with distinct flavors and characteristics.

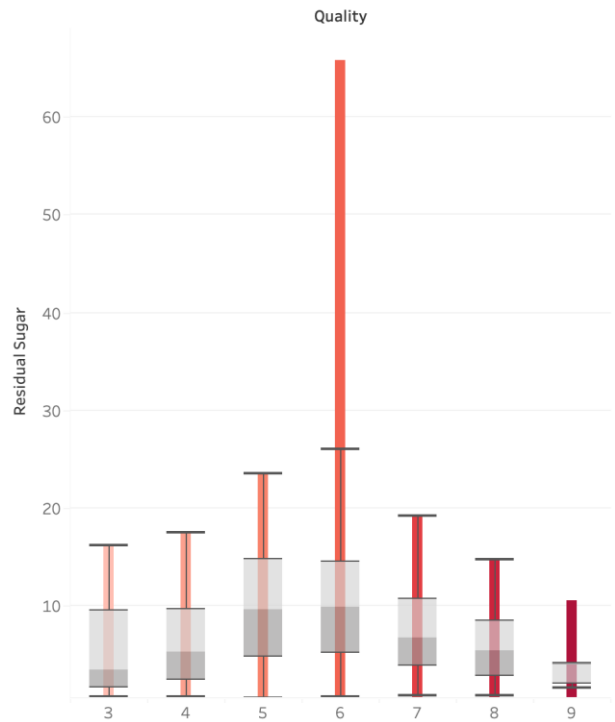
Distinguishing Quality Factors



Wine with a higher alcohol concentration tends to be of better quality. On average, taste tests have found that the ideal alcohol content averages around 13%.



Wines with a low pH will have a tart, fresh flavor, while those with a higher pH are more prone to bacterial growth. Most wines have a pH of 3 or 4, with white wines preferring a range of 3.0 to 3.4 and red wines preferring a range of 3.3 to 3.6.

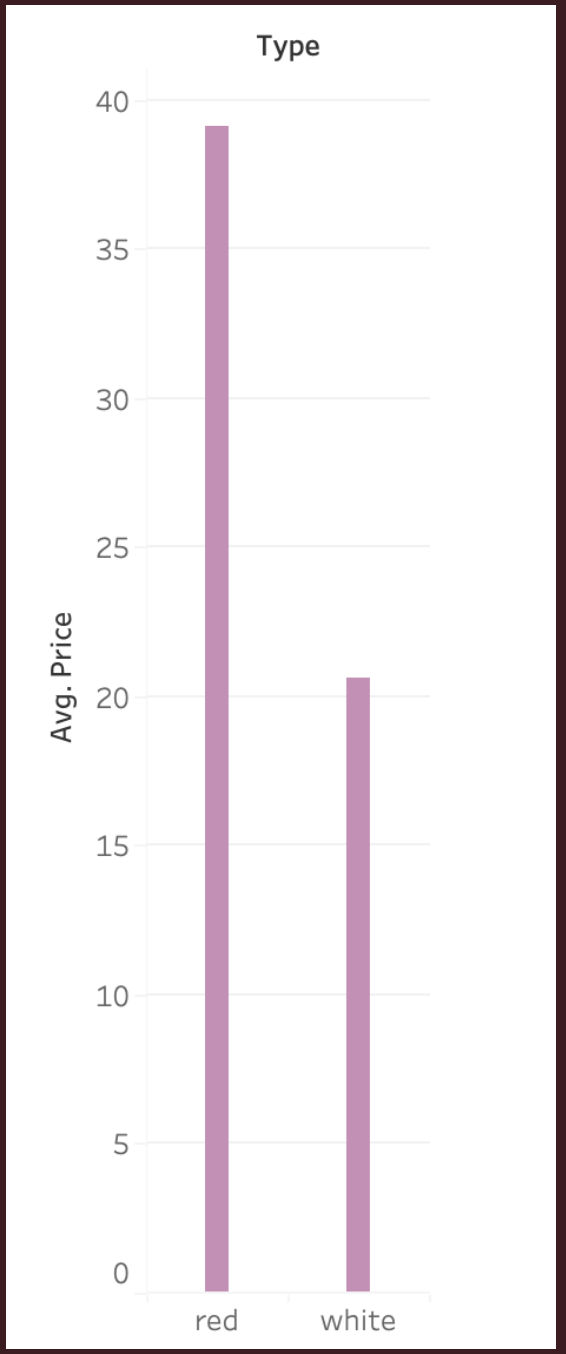


Residual sugar is the amount of sugar that remains in a wine after fermentation is complete. It can greatly impact a wine's taste, texture, and overall style. Wines with higher levels of residual sugar tend to be sweeter, while those with lower levels tend to be drier.

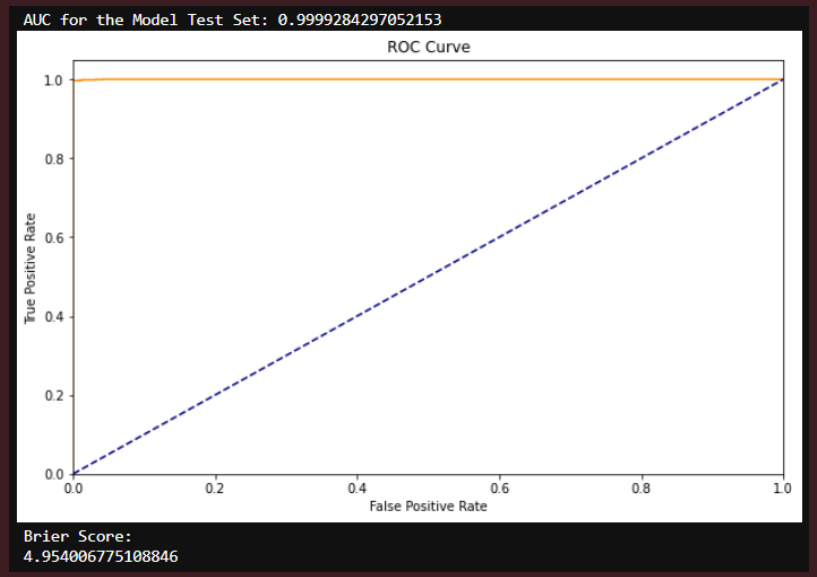
Price by Type

The four main types of wine are red, white, rosé, and sparkling. The two we chose to compare are red and white.

On average, red wine tends to be more expensive than white wine. The price of red wine is typically around $39, whereas white wine is typically around $20. Red wine is more expensive due to the production process for red wine being more labor-intensive and time-consuming than white wine.



Machine Learning Models



Type Roc Curve

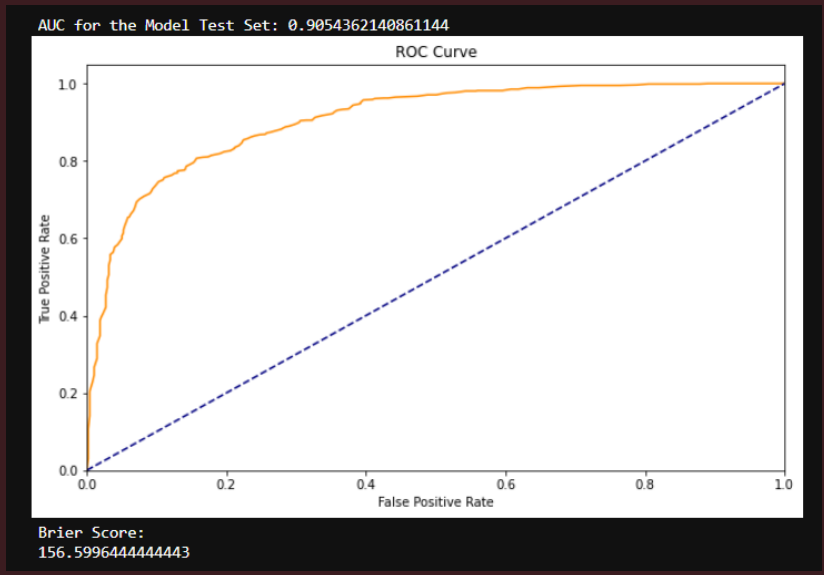
XGBoost Model

Our wine type classification model is most likely overfitted. We saved testing for overfitting for future work.

Quality Roc Curve

RandomForest Model

Our quality classification model is 82% percent accurate. Originally all the models had accuracy scores less than 75%, but after adding oversample = SMOTE() code the models improved.



Future Work & Limitations

There were certain limitations in our datasets and models. Due to the size of our Wine Rating & Price dataset, we had to limit it to only include ratings of 4-star and above, which could lead to unreliable conclusions regarding the regions with the most wine production. The Wine Quality Data Set (Red & White Wine) mainly consisted of wines rated at 5 or 6 quality ratings, which made creating a model that could predict multiple classifications unfeasible, as the model would have been skewed towards one classification most of the time. The accuracy score of our machine learning model for predicting wine quality did not meet our expectation of 90% and above, so the results of our quality prediction model may not be entirely trustworthy. To address this in future work, we intend to develop another machine learning model that utilizes location data, specifically longitude and latitude, to predict wine quality based on the region of origin. To achieve this, we would need to merge the datasets into a single CSV.

Conclusion

Our project was primarily successful in that we could answer the research questions we initially set out to answer. Although our quality prediction model was not as precise as we had hoped, we still consider the project to be a success as we achieved our objective of enabling users to utilize our model on our website to forecast the type and quality of the wine based on the given dataset features. We are confident that users will also be able to better appreciate the quality of their wine by using our analysis.


Thank You