BAX 452 Machine Learning

Uncorking Quality: A Data Driven Approach

Machine Learning Team

Sakshi Arya, Stuti Shekhar, Vivienne Xiang

Word count: 1,835

Table of Content

1. Executive Summary

The project aims to delve on ***"How chemical composition determines the Type and Quality level of different types of wine"***. This is done through the analysis of two comprehensive datasets consisting of data separately for white and red wine. Several advanced algorithms were applied including Decision Trees for classifying wine type and quality level, Random Forests for observing important features, and K-means Clustering to deep dive into observing the behavior of the model along with noting the accuracy.

The models yielded a **98.9%** of accuracy for predicting wine types through a logistic regression by taking in consideration all the chemical composition of wines. Decision trees achieved an accuracy of **93%** for classifying quality levels, and **96.1%** for wine type.

The search through data also provided valuable insights, unveiling the significant influence of specific chemical features. Random forest helped in determining Total Sulfur Dioxide as a pivotal factor in determining both wine type and quality level. Ingredients such as chlorides and Residual Sugar contributed majorly in describing the alcohol content.

2. Background

The wine market is highly competitive, with customers looking for more quality and variety in their selections. Understanding the chemical composition of wine is a significant step towards achieving consistent quality, optimized production which could potentially lead to creating wines tailored to customer preferences.

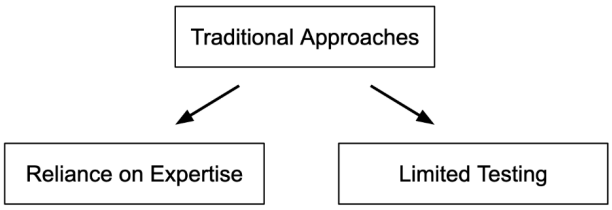
Vinho Verde, a Portuguese wine known for its tart acidity, low alcohol content, and subtle fizz, is the ideal summertime refreshment. The majority of the wines in this collection are white, making up 85% of the total; the remaining wines are either red or rosé. The dataset for analysis is obtained from UC Irvine

Machine Learning Repository, having a total of 4898 instances and 11 features modeled on physicochemical tests..

Owing to the wide range of wines in the current market, makers need to constantly innovate and set themselves apart from the competition in order to survive and thrive. This can be done by leveraging the **quality level analysis** which provides insights to the ingredients contributing to the caliber. Understanding the type along with this could help the makers develop distinctive products satisfying consumer preferences and outperform those of their competitors' in terms of quality.

3. Traditional Problem-Solving Approaches

Traditionally, winemakers relied on the expert knowledge of sommeliers and their experience. The accumulated knowledge on judgment of quality, often through taste tests, incorporates factors like aroma, taste and mouthfeel. A lot of subjectivity involved thorough but limited manual testing panels and lab testing.



However, due to the advancement in technology, data driven techniques could be incorporated in winemaking procedures. Wineries are able to conduct effective chemical analyses by utilizing large dataset and ML algorithms like decision trees and random forest. Because it improves the quality and streamlines the process, this method fits the business model and eventually makes the winery more competitive in the marketplace.

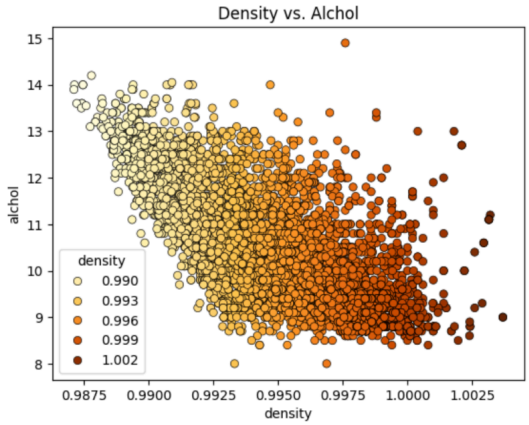
4. Analyses:

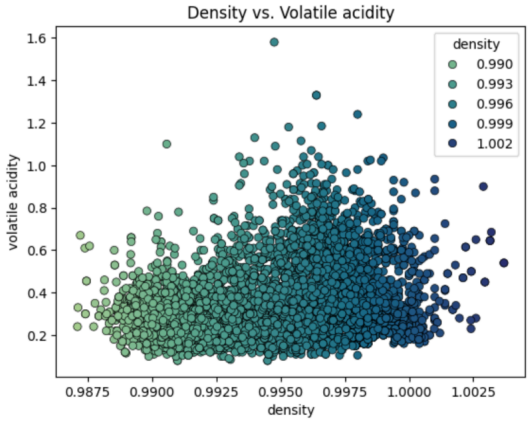
4.1 Data Exploration and Preprocessing

The data structure was analyzed along with exploring the statistical summaries. Two comprehensive datasets, having chemical composition of red and white wine, were combined by adding a type variable to distinguish between them. Additionally, a quality level was created by dividing the qualities into 3 bins or levels.

Columns	Description
Fixed Acidity	Amount of non-volatile acids present, mainly tartaric acid and malic acid. Contributes to sourness and tartness
Volatile Acidity	Amount of easily evaporated acids, mainly acetic acid
Citric Acid	A key ingredient that contributes to tartness and freshness
Residual Sugar	Amount of unfermented sugar remaining after fermentation. Affects sweetness and mouthfeel
Chlorides	Level of chloride salts, which can influence acidity and bitterness
Free Sulfur Dioxide	Amount of unbound sulfur dioxide, used as antimicrobial agent in winemaking process
Total Sulfur Oxide	Aggregated amount of free and bounded sulfur dioxide
Density	Mass per unit volume of the wine, usually influenced by alcohol and sugar content
pH	Measure of the acidity or alkalinity of the wine on a scale of 0 (highly acidic) to 14 (highly basic).
Sulphates	Level of sulfate salts, which can contribute to a "chalky" taste
Alcohol	Percentage of ethanol by volume, affecting warmth and sweetness perception
Quality	Score of 1- 9 indicating the perceived overall quality of the wine
Quality Level	Quality divided into 3 levels - [3,4]; [5,6,7]; [8,9]

Visualizations: Visualizations were plotted to check the relationship between the variables by selecting the columns accordingly. The below graph suggests a negative linear relationship between density and alcohol after handling outliers, however it does not suggest any clear relationship between density and volatile acidity.





4.2 Data Driven Analyses

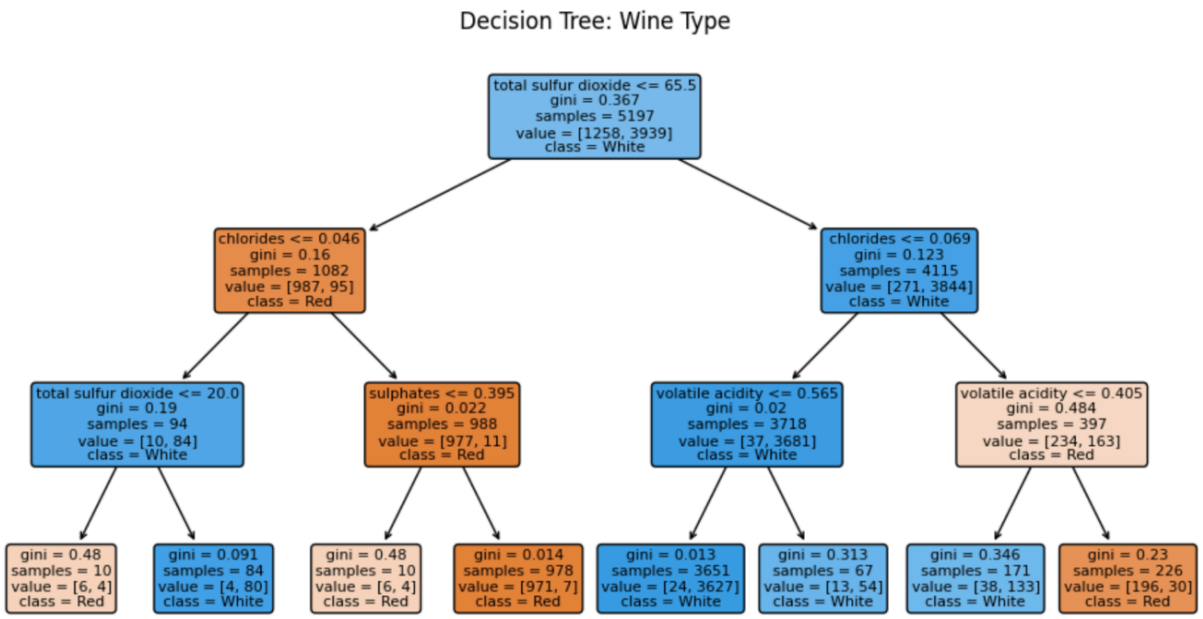
4.2.1 Logistic Regression

A logistic regression was run to predict the binary outcome for wine type - 'Red' or 'White' . The model was run with all the independent variables containing the chemical composition of the wine along with the quality. These were then used to predict the 'Type' variable.

An accuracy of around **98.9%** was obtained, which suggests that the model correctly predicts 98.9 % of the instances referring to a good model overall.

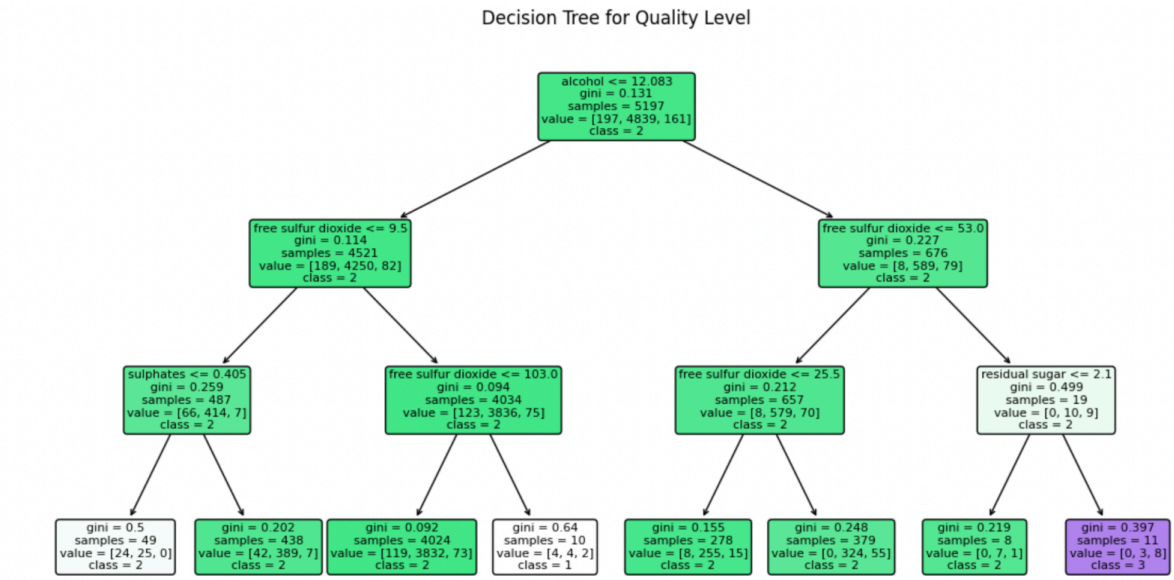
4.2.2 Decision Tree

A decision tree was plotted to predict wine type, identifying total sulfur dioxide, chlorides, and volatile acidity as important features. The root node has a total sulfur dioxide content ≤ 65.5, meaning it is the best feature to distinguish red wine and white wine. An accuracy of **96.3%** was achieved in classifying.



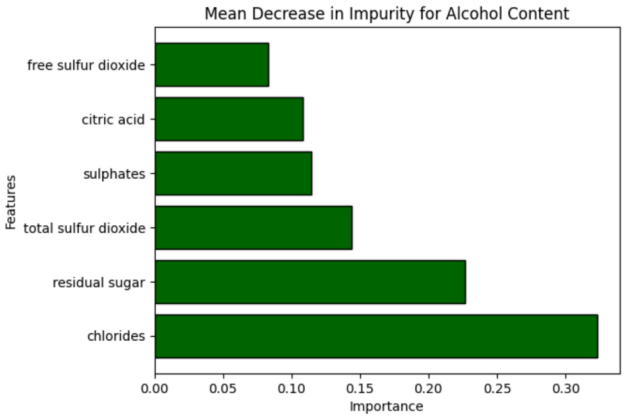
Pruning: The decision tree was pruned while fitting the dataset with a *max_depth = 3*. This was done, since the classified leaf nodes had samples < 10 and the tree obtained was very scarce. Additionally, a parameter of *min_samples_leaf = 10* was added, which constrains the nodes to split so that the leaf nodes at least have samples of size 10. This helps in forming more confident intervals leading to better accuracy.

Another decision tree with an accuracy of **93%** was constructed to predict the quality level, highlighting the importance of alcohol, free sulfur dioxide, and residual sugar. The root node is the alcohol content with a level of 12.083, making it the best feature to classify wine qualities. A constraint of *max_depth = 3* was applied to prune the tree to avoid overfitting the data with branches having little significance.



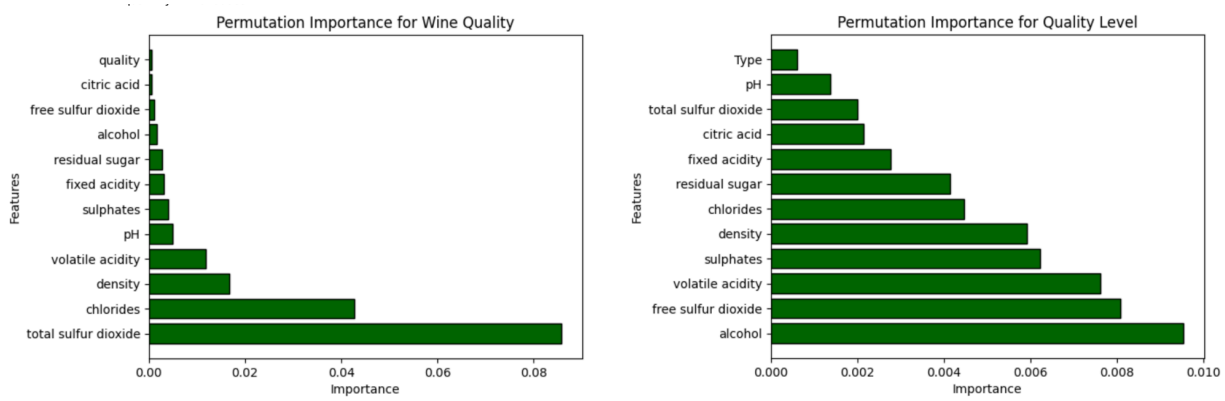
4.2.3 Random Forest

Random forest is a set of decision trees implemented using bootstrap with replacement. The trees were fitted with CART and then the average value was estimated. The nodes are selected by checking the highest decrease in deviation at each level. For selecting the most important features from Random Forest for the question: *"How chemical Ingredients contribute to alcohol content"*, mean decrease in impurity was plotted with Importance and Features.



The model suggests that Chlorides, Residual sugar, and Total Sulfur dioxide are the top 3 chemicals that contribute to alcohol content.

Additionally, Permutation Importance was applied to check how the deviance changes after shuffling columns for predicting wine Type and quality levels.

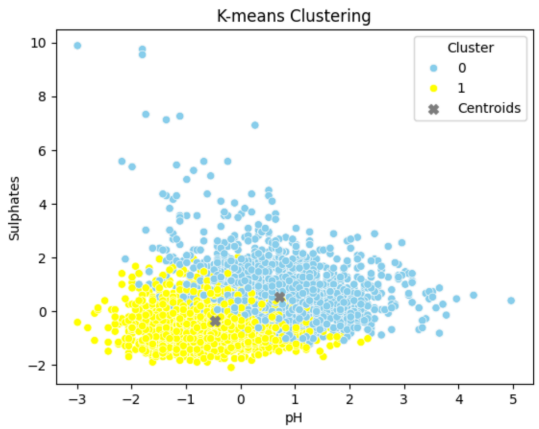


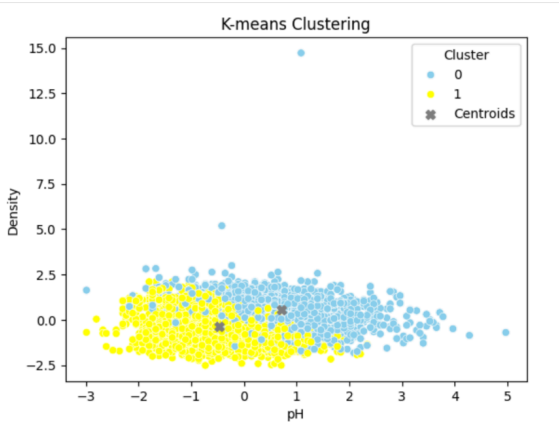
The result indicates that the most important feature for classifying Red and White wines is total sulfur dioxide, followed by chlorides, and density. For quality level , alcohol content plays a key role . The plot shows how the feature scores have a significant decline in Wine Quality among each other whereas have less decline in Quality level. An accuracy of **99.4%** was achieved for wine type and **94.6%** for quality level.

Random Forest helps in better estimation compared to decision trees since there is no need of handling overfitting. The output is estimated by averaging the trees with the related noise becoming very small when compared to one single tree. Variable restriction while producing each subset also aids in helping the overfitting. Along with this, it does not require handling too many parameters.

4.2.4 K-means Clustering

To deep dive into how different chemical components define wine types, k-means clustering was applied. Clusters were plotted with pH, density and Sulphur due to the distinction they provide in Red and white wine.





The clusters in both the plots suggest **'1'** which is encoded as white wines having less pH and are more acidic. But, an interesting observation was seen with Red Wine moving towards more Sulfur content. This aligns with the generally pattern of white wines having a less pH level between 3.0 to 4.0.

4.2.5 KNN Classification

KNN Classification for quality level was checked with k as 3, due to the quality levels specified before while creating the term. An accuracy of **92.4%** was achieved while estimating the model.

5. Recommendations and Business Value:

For wine producers looking to improve the quality of their offerings, the information provided can be quite important. Recognizing the role of sulfur dioxide, chlorides, and sulfates in red and white wine classification enables winemakers to adjust these constituents to get desired color and taste profiles. With the information based on quality levels, wineries can improve their production processes and recipes to continually produce high-quality wines that satisfy customers. They can also personalize their marketing activities, by focusing on campaigns leading to enhanced impact, and understanding customer preferences for acidity levels in wines.

Maintaining precise control over the consistency of the product and alcohol content by keeping an eye on important factors like residual sugar and chlorides can improve consumer happiness and build brand

loyalty. This could also lead to consistent quality throughout batches in factories. Wineries can find patterns and maximize operational efficiency by analyzing data clusters based on density and pH. This could help them to enhance the controlled variables while producing. They can take advantage of the analysis, keep ahead of market trends, and move in the direction of long-term success by using strategic moves.

6. Summary and Conclusions:

The project successfully suggests a good prediction level for the type of wine given all the chemical compositions. It also points out how sulfur dioxide, chlorides are important to classify the Red and White Wine. Decision Trees were also used to determine the optimal nodes for various quality levels, notably alcohol and free sulfur dioxide. Knowing the results of the classification and regression model included in the supervised algorithm, the model deep dived into chemical ingredients contributing to the alcohol content. The feature importance suggested the independent variables such as 'Chlorides' and 'Residual Sugar' are of the highest variable importance. After examining every characteristic, the most important feature for each type of wine was found to be "total sulfur dioxide," which was once more determined to be significant utilizing permutation importance, and alcohol, which was determined to be crucial for "quality level".

Implementing the k-means method to cluster the data according to "pH," "density," and "sulfates" yielded an intriguing result. The clusters with two labels for type - 'Red' and 'White' were developed signifying white wines forming in the left of the x-axis having less acidic nature. Although, there was a noticeable difference in the way the sulfates were grouped, with red wines having greater sulfate contents.

Future Work: Further research could explore additional features, incorporate expert knowledge, and validate findings on larger datasets with more attributes.

References

Cortez,Paulo, Cerdeira,A., Almeida,F., Matos,T., and Reis,J.. (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T.

MasterClass, (2021). Understanding the Many Wines From Portugal's Prominent Wine Region. https://www.masterclass.com/articles/what-is-vinho-verde

Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Abhishek Shah., (2022) Visualizing Data Using K-Means Clustering Unsupervised Machine Learning https://medium.com/@jwbtmf