Wine Quality Dataset Analysis

presented by Death Star Group 1

Margarita Aynagoz, Jonathan Hays, Jesse Parent, Blair Sonnen

Wine Quality is Subjective.

Right?

Problem: quality perception is often driven by

- ***Vintner heritage***
- ***Sensory analysis***
- ***Subjective opinions of sommeliers, etc.***
- ***Brand presence***

These are not data-driven inputs.

Can we modernize this (with ML)?

Our solution says yes:

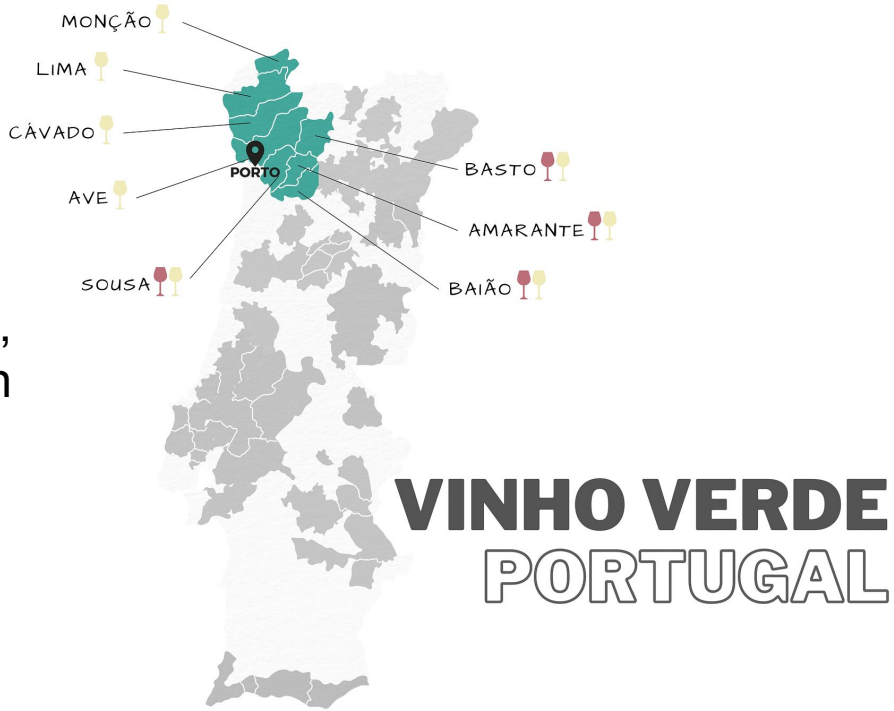
- Measurable composition (physicochemical features)
- Availability of chemical analysis techniques
- Repeatable and reproducible results
- Seemingly endless test subjects

These ARE data-driven inputs.

Good Data Can Provide Proof.

We can prove this using ML analysis:

- 6500-row dataset
- Red and White wine
- chemical properties and quality rating,
- sourced from a region in northwestern Portugal called Vinho Verde.
- Uses oenologist wine evaluations

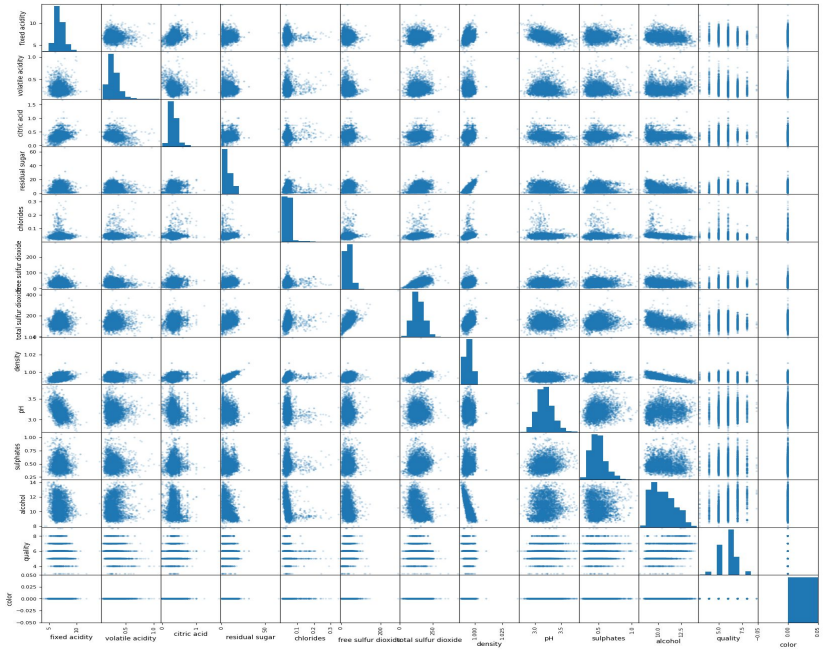


Wine Quality Dataset from the UCI Machine Learning Repository

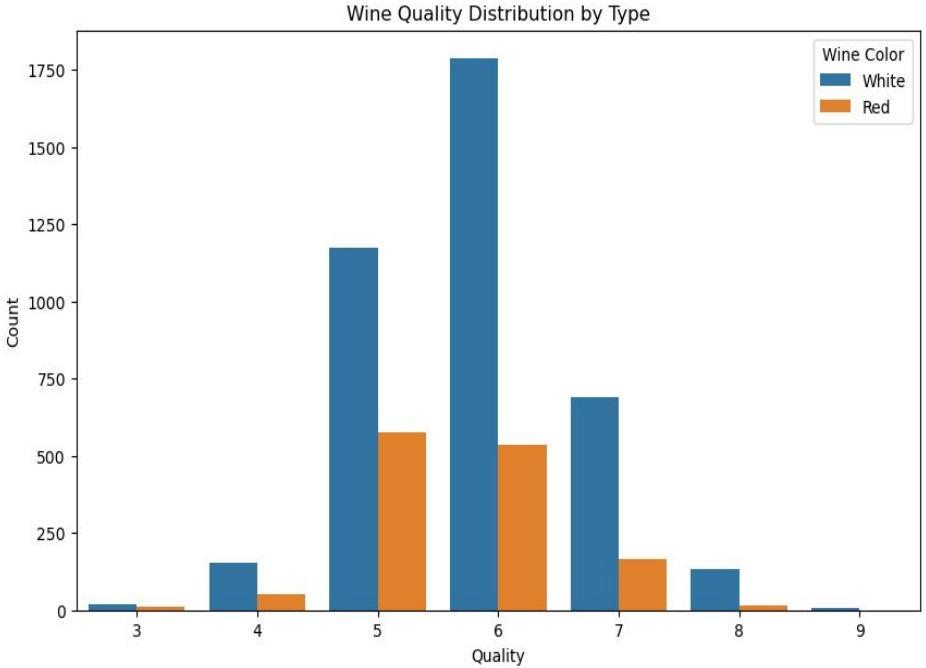
We envision a future where producers can adapt their methods based on analytical findings, resulting in improved wine quality and greater customer satisfaction.

Wine - A Symbol of Celebration

- A beverage that brings people together.
- Elevates ordinary moments to extraordinary experiences.



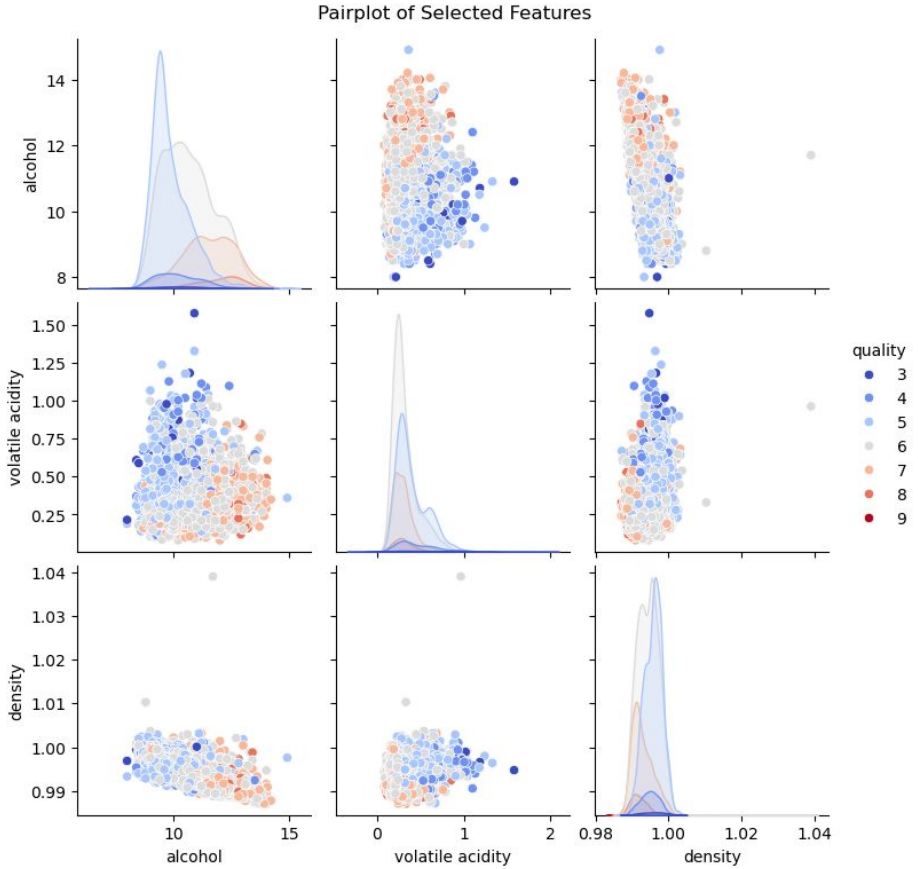
Visualizations of features and their relationships with the target variable **quality**.



- Histograms and boxplots to visualize feature distributions.

- Correlation matrices to identify relationships between features and overall quality.

- Create bins for the quality column.



Data Cleanup and Feature Selection:

- Drop the **“color”** column due to it not being relevant to quality.
- Drop the **“free sulfur dioxide”** column due to it being highly correlated with **“total sulfur dioxide”** column.
- Drop the **“chlorides”**, **“citric acid”**, **“fixed acidity“** columns because of the P-value did not significantly influence quality predictions.

Allow us to focus on the most impactful variables that truly contribute to the quality of wine.

○Approach: Jesse

Combine the white and red wine datasets with a new color column

Clean the data (drop nulls (0), de-duplicate (1,177), high correlation (1))

Visually analyze the data and look for relationships and shape of data

- Notice relationships
- Detect imbalanced output

Use Binning for quality into 0 (bad - 0-5) and 1 (good - 6-10)

- Avoids heavily imbalanced data
- Higher accuracy

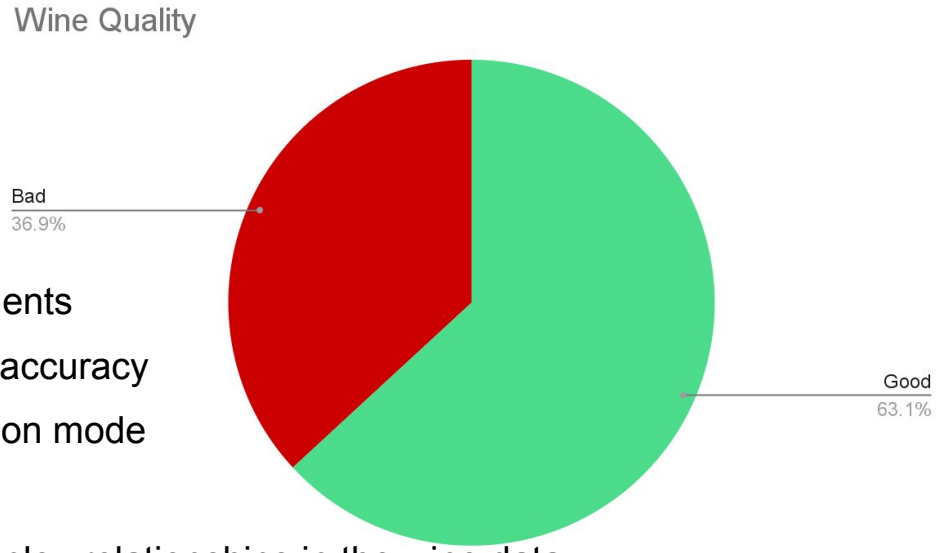
Analyze feature importance (p-values - Binary Logistic Regression) and drop less important features

- P-Values greater than 0.05 were dropped (chlorides: 0.073,fixed acidity: 0.58, citric acid: 0.65)
- We did not find a benefit from dropping features in model analysis

Scale Data

- Not necessary for all models but benefits Logistic Regression (slight benefit found)
- RandomForest handles unscaled data well, but is not adversely impacted by scaling

○Approach: Jesse



Rebalance data using SMOTE and SMOTEENN

- More good quality wine results than bad
- Random Forest showed no accuracy improvements
- Logistic Regression showed negative effect on accuracy

Model Selection - Binary output (quality) = classification mode

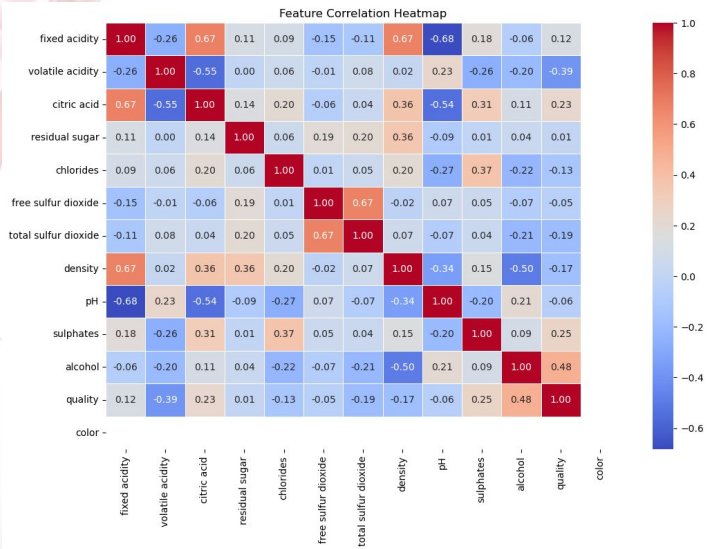
- Random Forest (Accuracy 76-77%)
  - Best for high accuracy and capturing complex relationships in the wine data
- Logistic Regression (Accuracy 76-77%)
  - Best for understanding which factors (e.g., acidity, alcohol) strongly impact wine quality

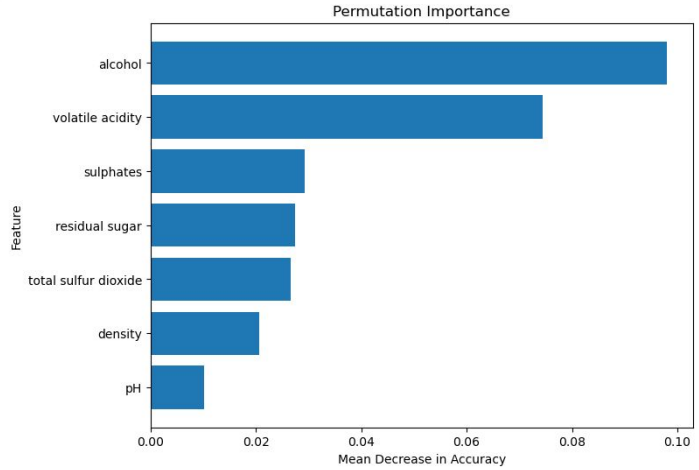
Hyperparameter Optimization

- Random Forest showed 1% additional accuracy increase (78%)
- Logistic Regression showed no significant improvement

Conclusions: Jonathan

- Alcohol is the strongest predictor of wine quality
- Volatile acidity and density are strong negative indicators
- Some features (e.g., total sulfur dioxide, pH, color) have minimal impact
- Simplifying models by removing low-impact features improves efficiency
- Correlation and feature importance metrics aligned well





What could have been next?

- Analyzing our results as compared to the 2009 paper from Cortez et al.
- Performing additional model runs on SVM with hyperparameter tuning.
- Running these models on Red and White datasets separately.
- Finding more recent data from another region
  - Or a large volume vintner
- Possibly merging crowdsourced ratings into dataset if possible