Modeling Wine Quality: Investigating the Role of Alcohol Content

Authors:
Patrycja Szostak,
Oliver Tischer,
Luis Dlugos

Research Question:

Does higher alcohol content lead to better wine quality?

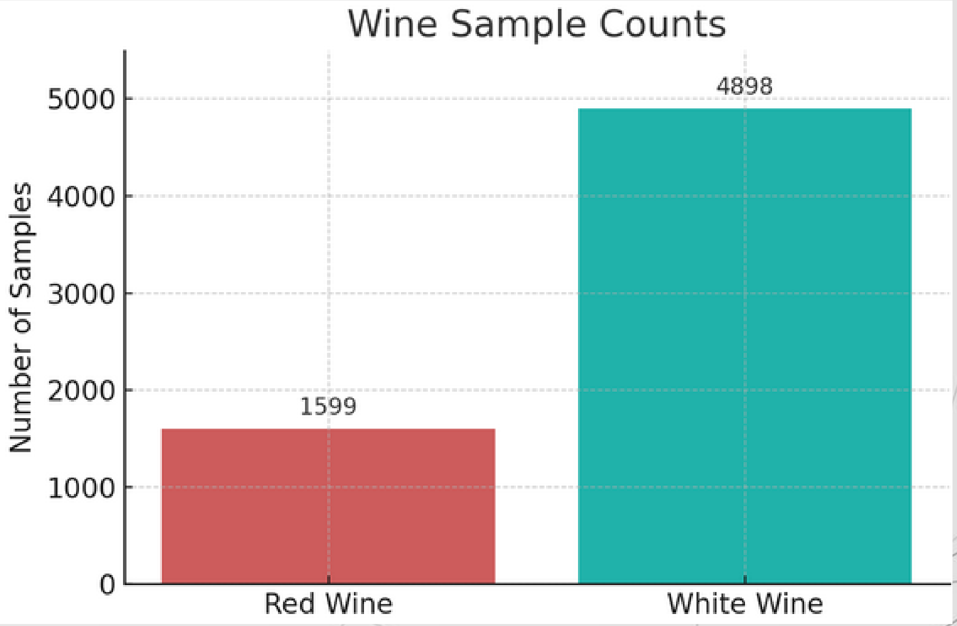
01 INTRODUCTION

Introduction & Dataset Overview

- **Source:** UCL Machine Learning Repository
- **2 Datasets:** Red wine (winequality-red.csv) and White wine (winequality-white.csv)
- **Samples:**

Red wine: 1599 entries
White wine: 4898 entries

- **Features:** 11 physicochemical variables
- **Target variable:** quality
- No missing values



Patrycja Szostak

02

Chemistry behind wine features.

Acidity and Taste Balance

Features that shape the wine's freshness, structure and drinkability.

- **Fixed Acidity -** primary tartaric and malic acid.

In red wine: sharpness, freshness; balances fruitiness. Harsh taste if too much.
In white wine: crispness, as white more acid-driven.

- **Volatile Acidity -** mostly **acetic acid** (vinegar component).

In red wine: High levels indicate spoilage. Low levels are normal.
In white wine: Even small increases are noticeable and undesirable.

- **Citric Acid -** used to add acidity, can be an additive.

In red wine: less common.
In white wine: Sometimes added to enhance freshness in whites. Can slightly improve taste.

- **Residual Sugar -** unfermented sugar (glucose/fructose) left after fermentation.

In red wine: various values, gives different sweetness.
In white wine: Slightly higher than reds on average. Important for balance.

- **pH** - Measures **acidity strength.** Lower pH = more acidic

In red wine: around 3.4–3.6. Affects stability, taste, and colour tone.
In white wine: Lower pH (3.0–3.3) more common. Helps preserve freshness.

Patrycja Szostak

Stability and Preservation

Impact shelf life and oxidation.

- **Chlorides** - represent salt content, mainly sodium chloride.

In red wine: rare in red wines, if present salty and metallic notes.
In white wine: more sensitive to presence of chlorides (may come from storage or winemaking water).

- **Free Sulfur Dioxide** - antioxidant and antimicrobial agent.

In red wine: preserves freshness and colour, used sparingly.
In white wine: critical, prevents oxidation and browning.

- **Total Sulfur Dioxide** – all bound and free SO2 in the wine.

In red wine: lower, as tannins and anthocyanins give natural protection.
In white wine: significantly higher, essential. Keep freshness and stability.

- **Sulphates** – potassium metabisulfite. Added but also occurs naturally.

In red wine: aid in microbial control and shelf control. Moderate use.
In white wine: provide stability and freshness. Higher use.

Patrycja Szostak

Body, Strength & Overall Quality

Drive the wine's mouthfeel and scoring.

- **Density -** mass per volume (g/mL). Correlates with **alcohol** and **sugar content**.

In red wine: lower density usually means higher alcohol, higher density – may suggest sugar.
In white wine: same as in red.

- **Alcohol -** ethanol, produced during fermentation from sugar.

In red wine: higher alcohol enhances mouthfeel and perceived sweetness.
In white wine: similar effects, might overpower lighter wines.

- **Quality (target feature) - human sensory rating based on taste, aroma and balance. Given by professionals.**

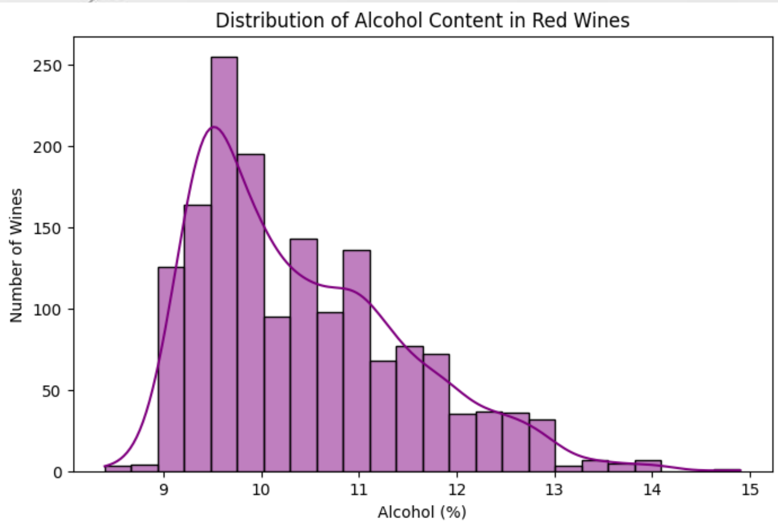
In red wine: more body, tannin and complexity expected (structure).
In white wine: Balance of freshness, fruit and acidity. Preferable clean and aromatic profile.

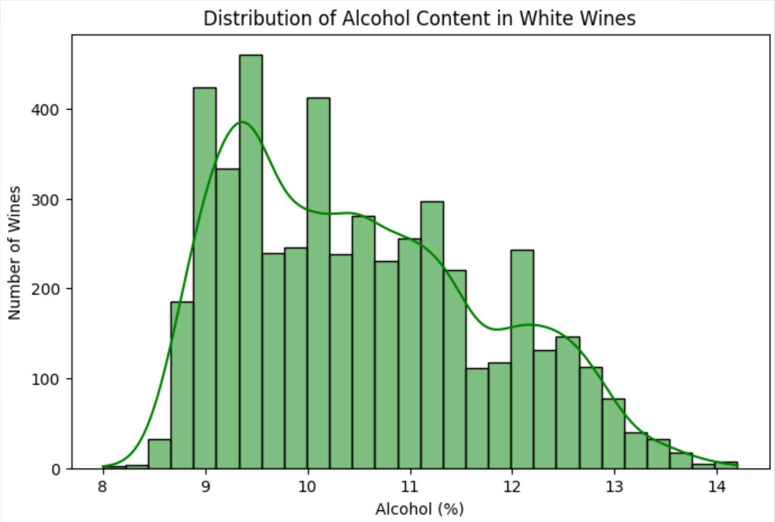
Patrycja Szostak

03

Exploratory Data Analysis (EDA): Getting to Know the Wines

Alcohol Distribution





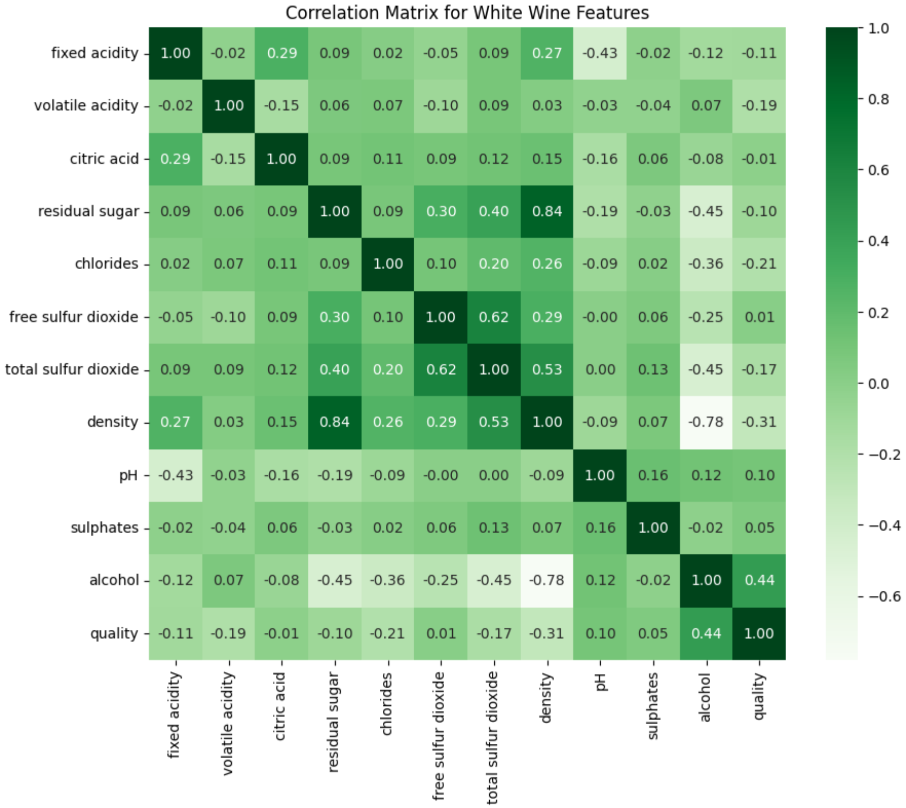
Average alcohol content: 10.42%
Alcohol range: 8.4% - 14.9%

Average alcohol content: 10.51%
Alcohol range: 8.0% - 14.02%

Patrycja Szostak

Feature Correlations

Which Features Drive Wine Quality?



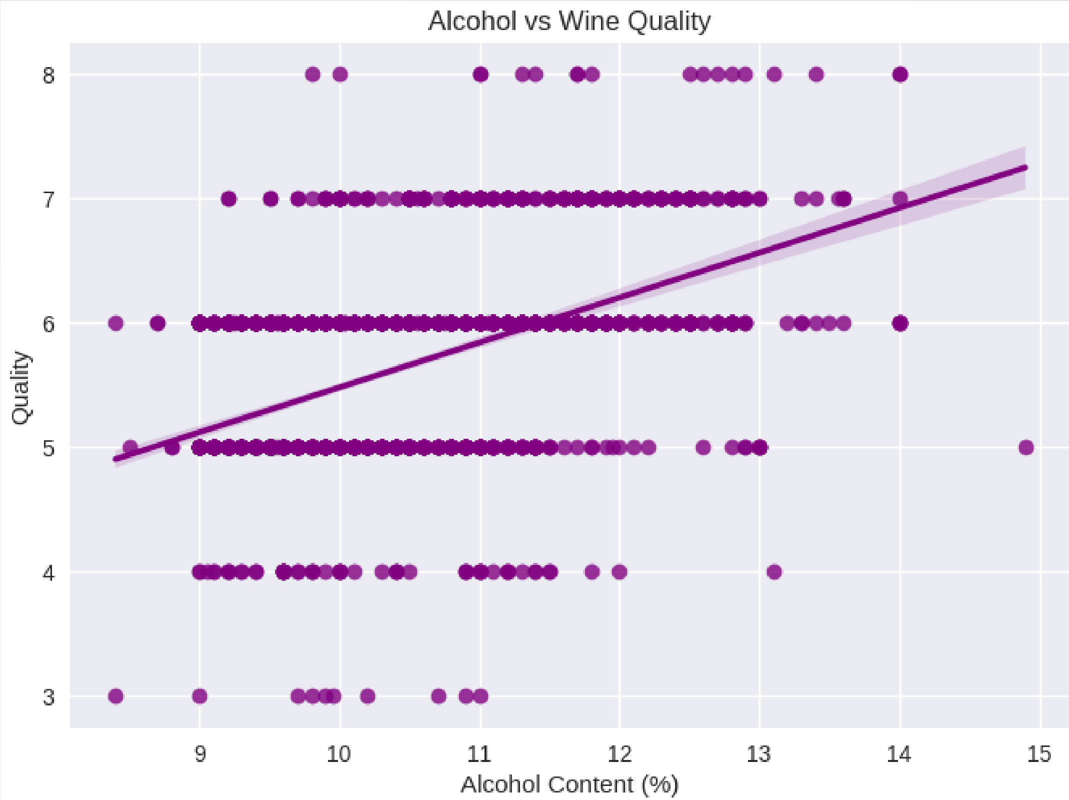
Patrycja Szostak

Feature Correlations

Which Features Drive Wine Quality?

Patrycja Szostak

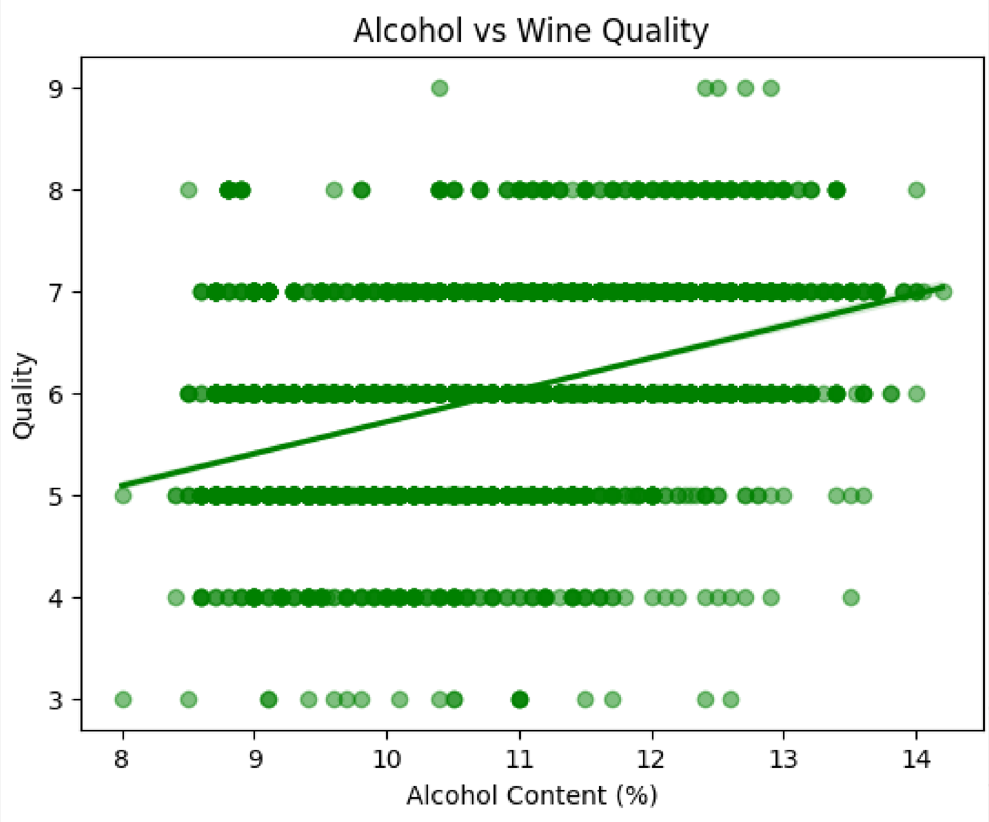
Pearson Correlation – red wine



Pearson Correlaton score: 0.4762

Patrycja Szostak

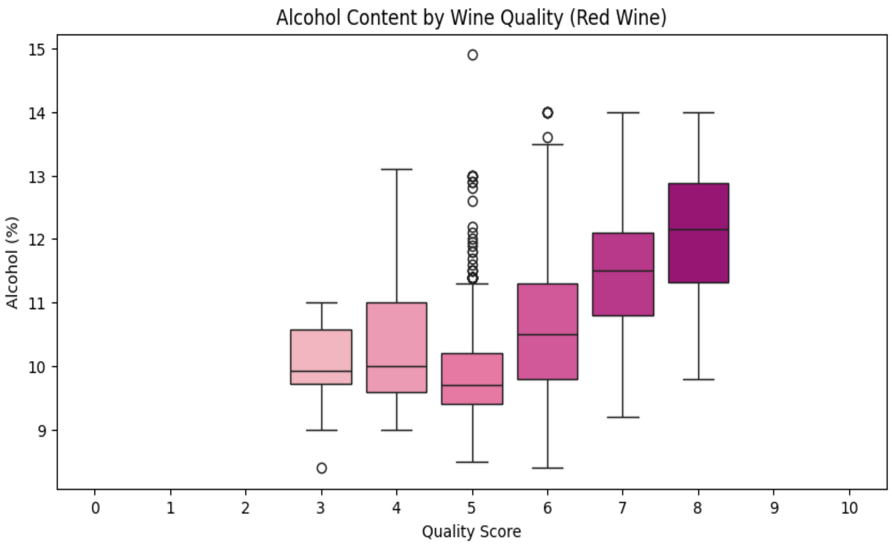
Pearson Correlation – white wine

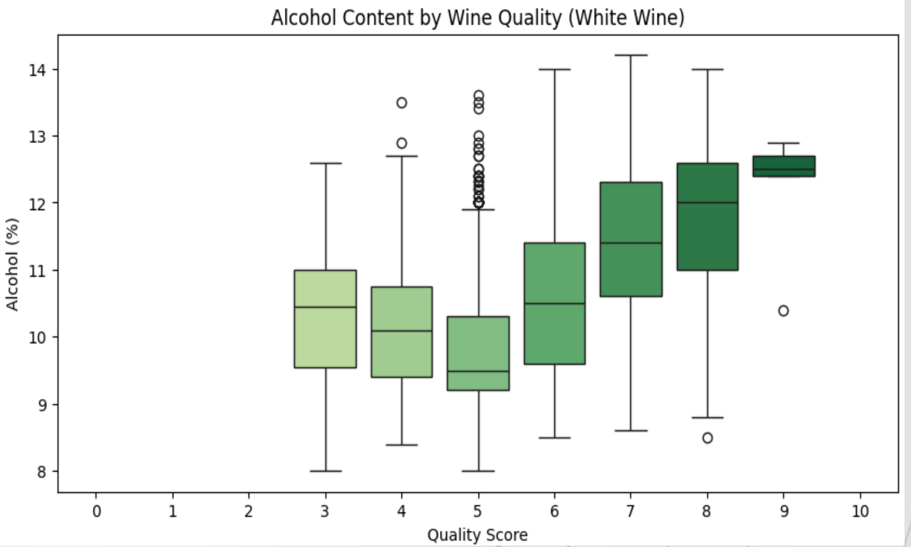


Pearson Correlaton score: 0.4356

Patrycja Szostak

Alcohol vs. Quality





Strong positive correlation: Both red and white wines show clear upward trends where higher alcohol content correlates with better quality scores, with quality 3-5 wines averaging ~10% alcohol versus quality 7-8 wines reaching ~12-13% alcohol.
Not perfectly linear relationship: The alcohol-quality relationship levels off at the highest quality wines, meaning very high alcohol doesn't always guarantee better quality (however upward trend can be observed, especially in red wine).
Wine type differences: Red wines demonstrate more dramatic alcohol increases with quality improvement, while white wines show more gradual progression and higher variability within quality categories.

Patrycja Szostak

Key Takeaways

- EDA reveals that **alcohol** is **likely** the most **influential predictor** of **quality**.
- Other features indicate influence on quality as well.
- EDA helps build intuition before applying machine learning methods.
- Wine type matters as red wines show slightly stronger alcohol-quality correlation (0.48) than white wines (0.44)
- **Not perfectly linear relationship observed** - quality improvement plateau at higher alcohol levels suggests complex interactions. This insight warns us: 'more alcohol = better wine' is an oversimplification — especially at higher quality levels – crucial information in terms of ML modeling.

Patrycja Szostak

04

Predictive Modeling Using Machine Learning

With a focus on alcohol content

Initial Modeling Phase – Red Wine

Decision Tree (all features)

- Baseline Model
- Confusion Matrix

Feature Importance Analysis

- Identify key predictive features
- Reduce dimensionality for simpler modeling

Decision Tree (Top Features)

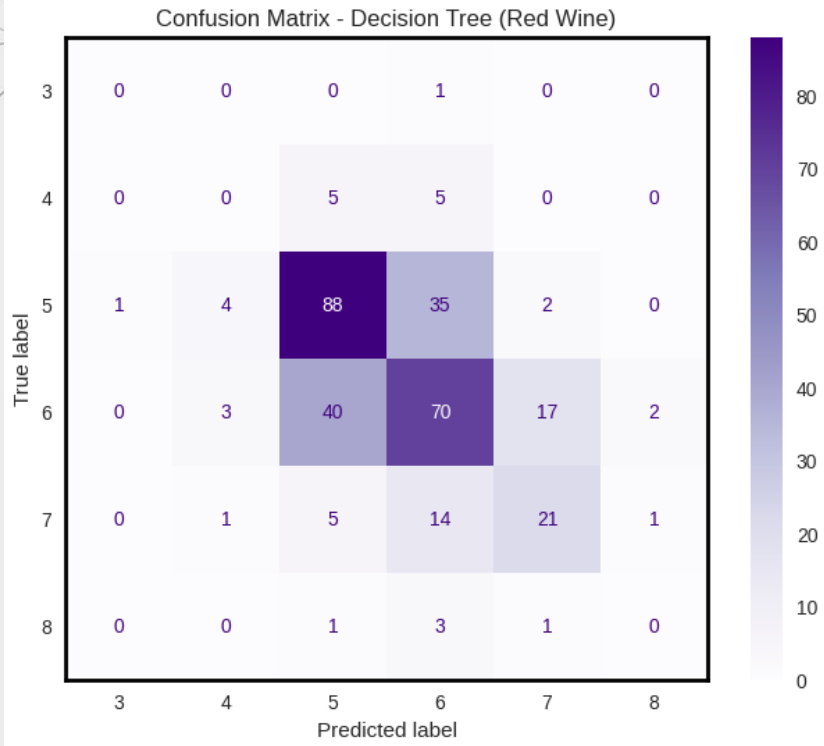
- Test model using top predictors only
- Compare performance with full-feature model

Random Forest (Top Features)

- Apply more advanced ML method
- Check improvement over single decision tree

Patrycja Szostak

Decision Tree - All Features



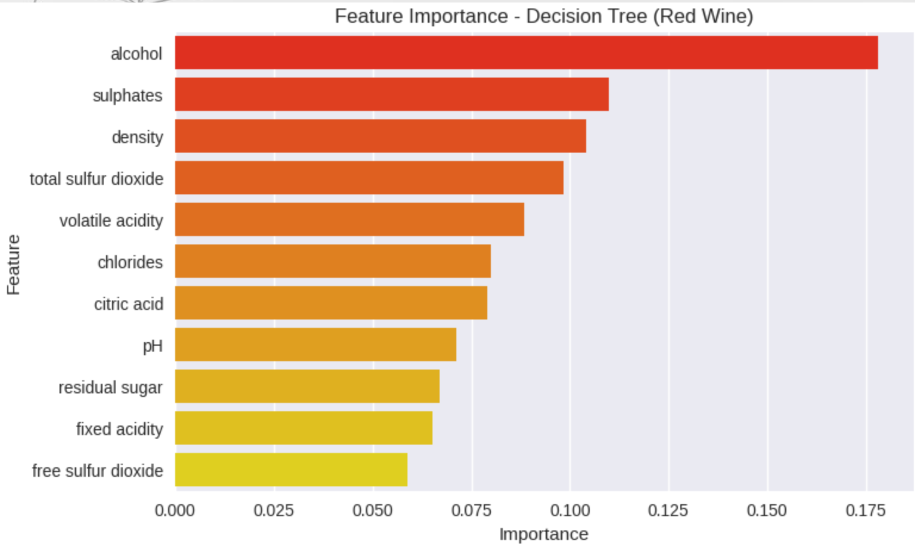
Accuracy: 0.56

Key insights:

- **Reasonable performance** (56% accuracy)
- **Better than random guessing** (16.7% for 6 classes: 3,4,5,6,7,8)
- **Best at middle range wines** – performs best on quality 5 (most common type) and struggles with rare – very low or very high quality wines
- **Most mistakes happen between neighboring scores** (5-6, 6-7)
- **Wine quality is difficult in predicting** and this model shows the complexity of the problem
- **Room for improvement** - this baseline shows our research question is worth pursuing

Patrycja Szostak

Feature Importance Analysis

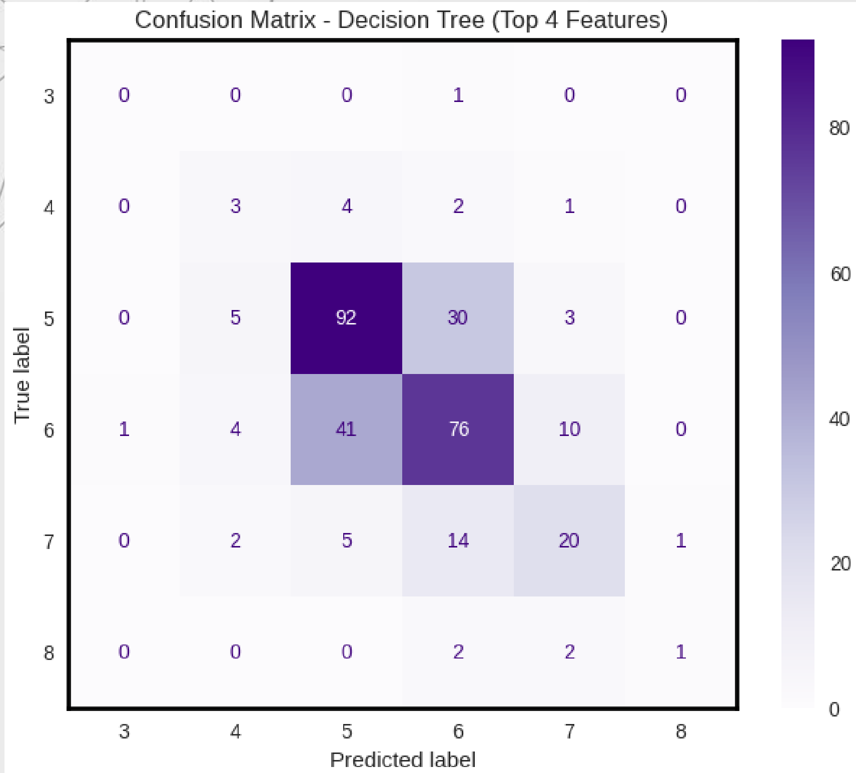


Rank	Feature	Importance
1	Alcohol	0.178035
2	Sulphates	0.109806
3	Density	0.103983
4	Total Sulfur Dioxide	0.098286
5	Volatile Acidity	0.088559
6	Chlorides	0.079945
7	Citric Acid	0.079002
8	pH	0.071373
9	Residual Sugar	0.066926
10	Fixed Acidity	0.065299
11	Free Sulfur Dioxide	0.058806

Feature importance shows which attributes the model relies on most — but alone, it doesn't guarantee predictive accuracy or answer our research question.

Patrycja Szostak

Decision Tree – Top Features



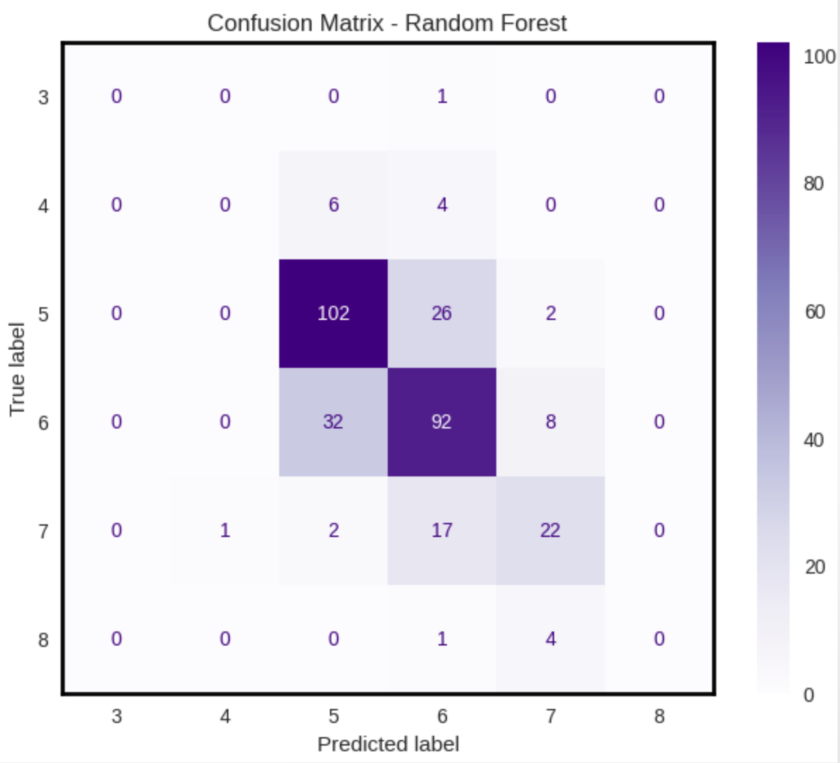
Accuracy: 0.60

Key Insights:

- **Best results with 4 top features** (alcohol, sulphates, density, total sulfur dioxide)
- **Improved performance** (60% accuracy vs. 56% with all features) - **top features outperform full model**
- **Most mistakes still between neighbouring scores** (5-6, 6-7) - pattern remains consistent
- **Efficiency gains** - better accuracy with fewer features (noise reduction)
- **Feature selection improves both simplicity and performance**

Patrycja Szostak

Random Forest – Top Features



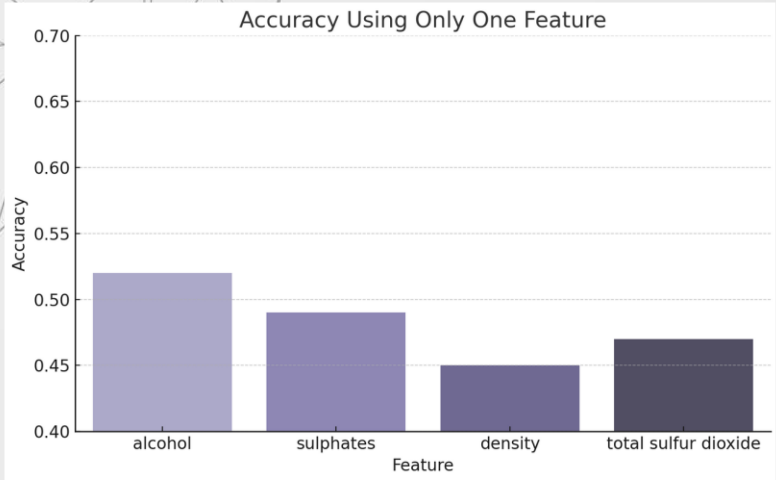
Accuracy: 0.68

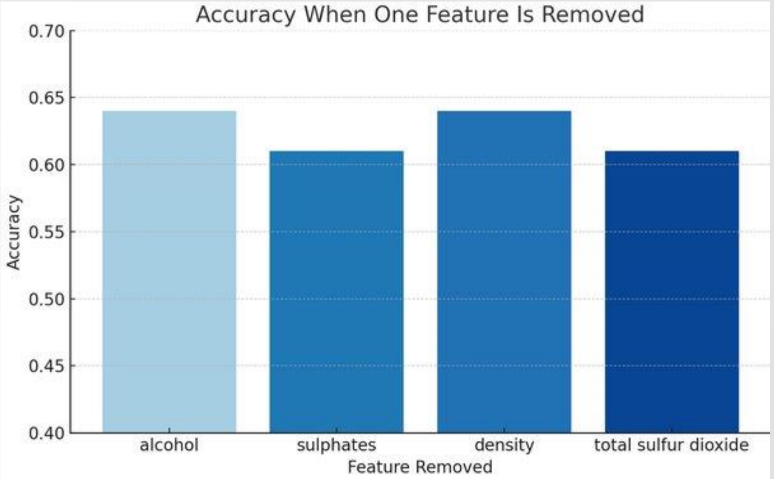
Key Insights:

- **Good performance** (68% accuracy)
- **Significant improvement** over Decision Tree (68% vs. 60%)
- **Strong performance on common wines** - quality 5 and quality
- **Most mistakes still between neighbouring scores** (5-6, 6-7) - pattern remains consistent across all models
- **Ensemble advantage confirmed** - multiple trees capture wine quality patterns better than single tree
- **Feature selection success validated** - top 4 features with Random Forest achieve optimal balance of simplicity and accuracy

Patrycja Szostak

Additional Ablation Tests – Red Wine





Only alcohol → 52%

Only sulphates → 48%

Only density → 45%

Only total sulfur dioxide → 47%

Without alcohol → 64%

Without sulphates → 61%

Without density → 64%

Without total sulfur dioxide → 61%

Patrycja Szostak

Initial Modeling Phase – White Wine

Decision Tree (all features)

- Baseline Model
- Confusion Matrix

Feature Importance Analysis

- Assess feature importance
- Keeping all features

Random Forest (All Features)

- Apply ensemble method
- Check improvement over single decision tree

Patrycja Szostak

Decision Tree – All Features

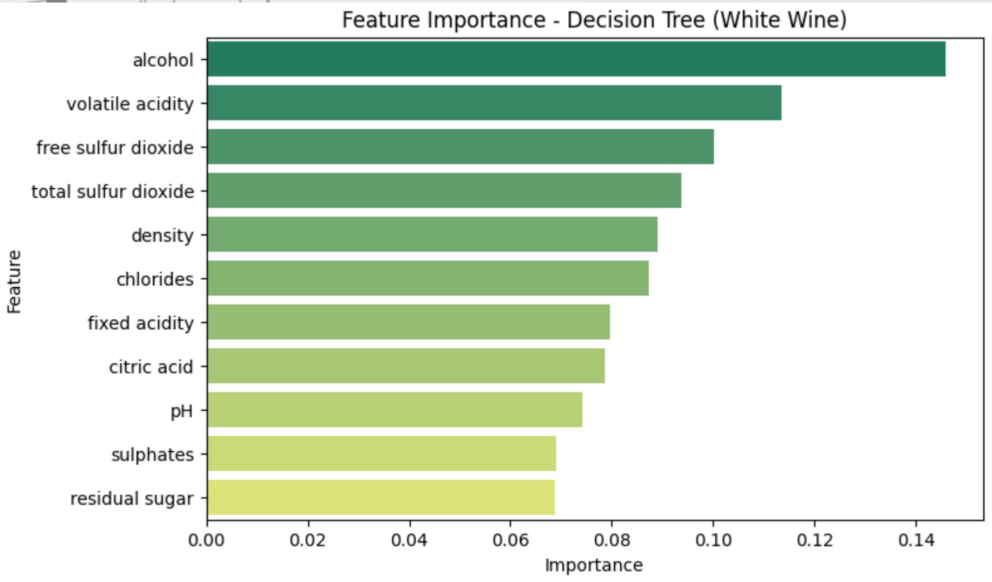
Accuracy: 0.61

Key Insights:

- **Good performance** (61% accuracy)
- **Better than random guessing** (14.3% for 7 classes: 3,4,5,6,7,8,9)
- **Best at middle range wines** - performs best on quality 5-6 (most common types) and struggles with rare quality wines
- **Most mistakes happen between neighboring scores** (5-6, 6-7)
- **Severe class imbalance** - extreme quality wines (3,4,8,9) rarely represented in predictions
- **Wine quality is difficult in predicting** and this model shows the complexity of the problem.
- **Room for improvement** - this baseline shows our research question is worth pursuing

Patrycja Szostak

Feature Importance Analysis

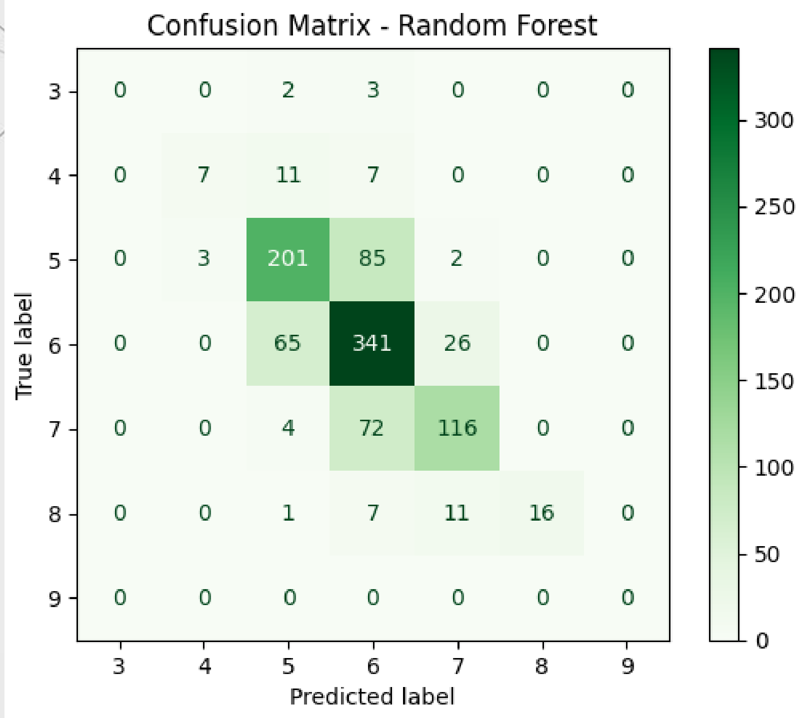


Rank	Feature	Importance
1	Alcohol	0.152
2	Volatile Acidity	0.119
3	Free Sulfur Dioxide	0.103
4	Total Sulfur Dioxide	0.099
5	Density	0.089
6	Chlorides	0.087
7	Fixed Acidity	0.082
8	Citric Acid	0.079
9	pH	0.073
10	Sulphates	0.067
11	Residual Sugar	0.065

Feature importance shows which attributes the model relies on most — but alone, it doesn't guarantee predictive accuracy or answer our research question.

Patrycja Szostak

Random Forest – All Features



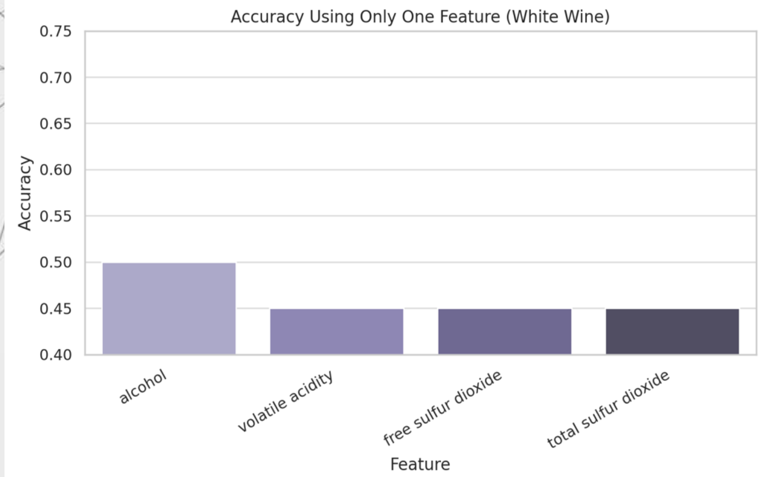
Accuracy: 0.69

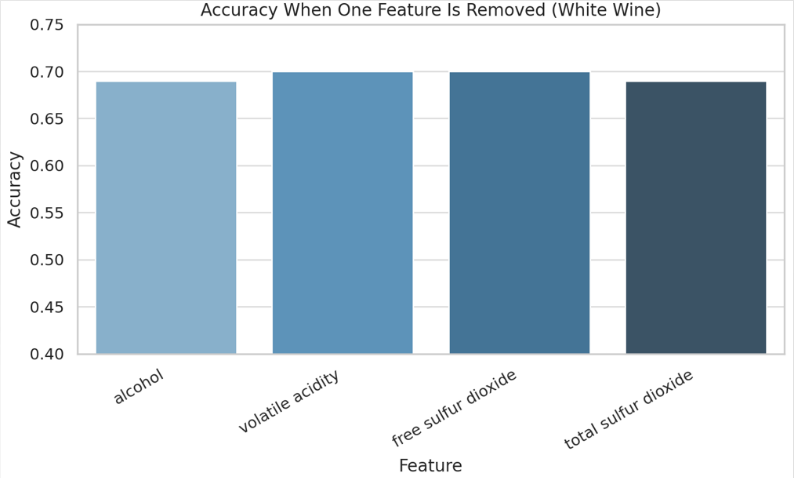
Key Insights:

- **Strong performance** (69% accuracy)
- **Better than baseline**
- **Excellent at middle range wines** - quality 5 and quality 6
- **Most mistakes happen between neighbouring scores** (5-6, 6-7)
- **Ensemble method advantage** - Random Forest captures complex wine quality patterns effectively
- **Severe class imbalance** - quality 3,4,8,9 poorly predicted
- Next step: class balancing techniques to boost minority class performance

Patrycja Szostak

Additional Ablation Tests





Only alcohol → 50%

Only volatile acidity → 45%

Only free sulfur dioxide → 45%

Only total sulfur dioxide → 45%

Without alcohol → 69%

Without volatile acidity → 70%

Without free sulfur dioxide → 70%

Without total sulfur dioxide → 69%

Key Conclusions – Ablation Tests & Feature Imbalance in Red & White Wine

- **Alcohol is strongest single predictor** - validates our research hypothesis across both wine types
- **Feature interactions matter more than individual features** - alcohol's impact depends on context with other chemical properties
- **Optimal feature selection varies by wine type** - red wines benefit from top 4 features, white wines improve by removing volatile acidity or free sulfur dioxide
- **While alcohol leads to better quality, it's not the whole picture. Presence of other important features is crucial for performance**
- **Models is robust** - no single feature is irreplaceable due to feature interdependencies
- **Severe dataset imbalance affects both types** - models heavily bias toward common quality scores (5-6). White wines showing more extreme imbalance
- **Performance concern** - decent accuracy may result from "smart guessing" dominant classes rather than true learning
- **Rare quality prediction fails** - extreme scores (3,4,8) consistently misclassified due to insufficient samples
- **Next step** - dataset balancing is needed for more reliable evaluation

Patrycja Szostak

05

Addressing Class Imbalance in Wine Quality Data

The Class Imbalance Problem in Red Wine

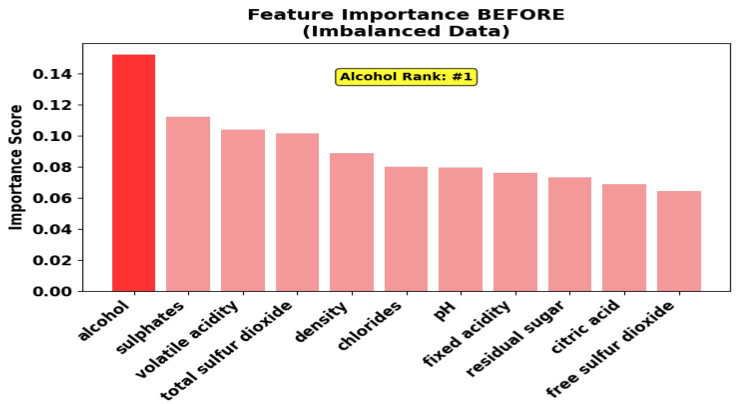
SEVERE CLASS IMBALANCE DETECTED:

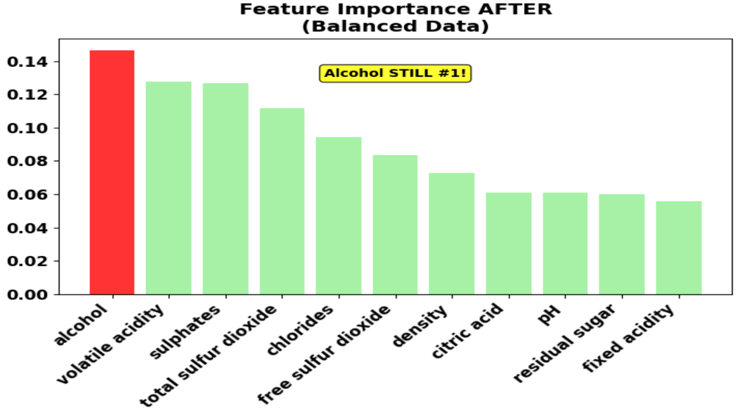
- \- Quality 5: 681 samples (42.6%)
- \- Quality 6: 638 samples (39.9%)
- \- Quality 3: 10 samples (0.6%) ← CRITICAL PROBLEM
- \- Quality 8: 18 samples (1.1%)

IMPACT ON RESEARCH:

- \- Models biased toward common wines (5-6)
- \- Cannot reliably evaluate rare excellent/poor wines
- \- Research question compromised - alcohol's role unclear for ALL quality levels

SMOTE Solution & Validated Results for Red Wine





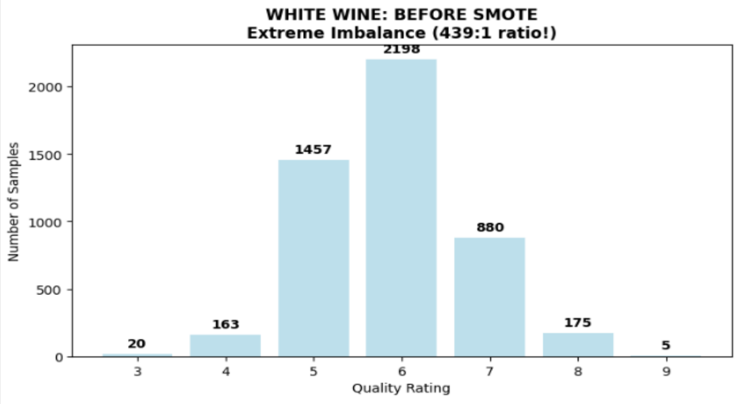
SMOTE RESULTS - RED WINE:

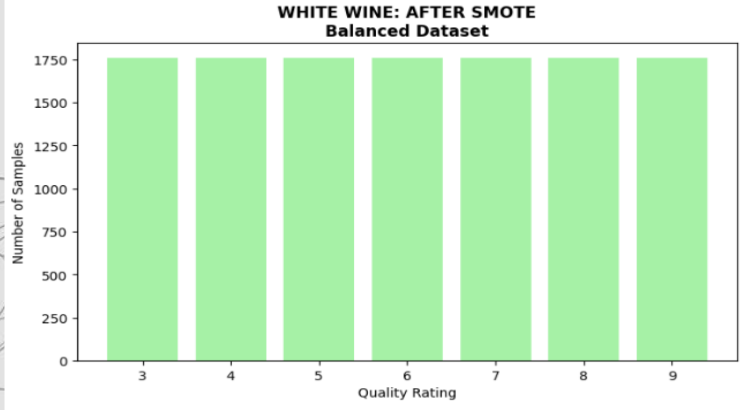
RESEARCH QUESTION CONFIRMED: Higher alcohol → better quality

KEY EVIDENCE:
- Alcohol remains #1 predictor (14.6% importance)
- Correlation unchanged: 0.476 (strongest)
- Quality 8 wines: 12.4% alcohol
- Quality 3 wines: 9.9% alcohol
- Difference: +2.5% alcohol for best wines

MODEL PERFORMANCE:
- Accuracy: 68% → 63% (expected trade-off)
- Better minority class detection
- More reliable across ALL quality levels

The Class Imbalance Problem in White Wine



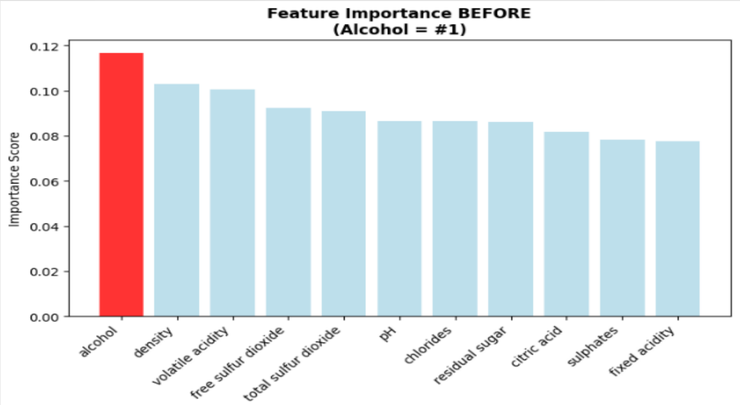


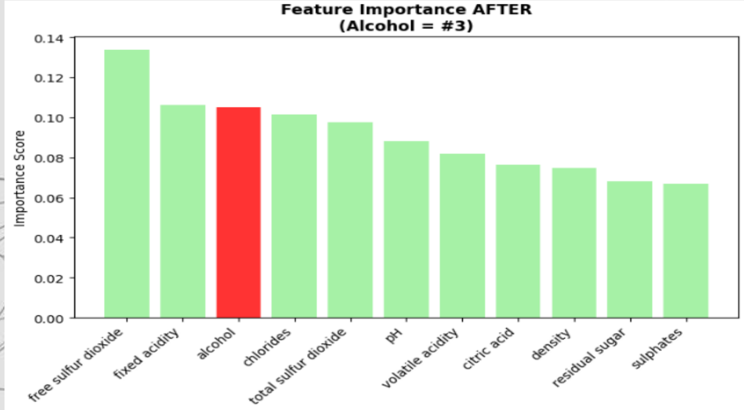
EXTREME CLASS IMBALANCE DETECTED:

- Quality 6: 2,198 samples (44.9%)
- Quality 5: 1,457 samples (29.7%)
- Quality 9: 5 samples (0.1%) ← CRITICAL PROBLEM
- Quality 3: 20 samples (0.4%)

IMPACT ON RESEARCH:
- Even worse than red wine (439:1 vs 68:1 ratio)
- 7 quality classes vs 6 in red wine
- Models extremely biased toward common wines (5-6)
- Cannot evaluate rare premium/poor wines
- Research question validity uncertain across ALL quality levels

SMOTE Solution & Validated Results for White Wine





SMOTE RESULTS - WHITE WINE:

RESEARCH QUESTION CONFIRMED: Higher alcohol → better quality BUT different pattern than red wine!

KEY EVIDENCE:
- Alcohol correlation: 0.436 (strong, but lower than red wine's 0.476)
- Alcohol ranking: #1 → #3 after SMOTE (reveals true importance)

WHITE WINE DIFFERS FROM RED WINE:
- Alcohol important but NOT dominant factor
- Free sulfur dioxide & fixed acidity more critical
- Preservation factors matter more in white wine

MODEL PERFORMANCE:
- Accuracy: 68% → 64% (expected trade-off)
- Better minority class detection
- More reliable across ALL 7 quality levels

Conclusions & Research Question

Does higher alcohol content lead to better wine quality?

Higher alcohol content is significantly associated with higher wine quality scores, especially in red wines. **However!** This is **not always the case**, and alcohol alone **does not fully explain** how wine quality is perceived.

Key Findings:

- Alcohol showed the **strongest correlation** with quality among all features.
- In **red wine**, the relationship was clearer and more consistent.
- In **white wine**, other features such as **free sulfur dioxide** and **acidity** had stronger influence after balancing.
- The relationship between alcohol and quality is **not perfectly linear** – beyond a certain point, higher alcohol doesn't always mean better quality.
- Machine learning models (Random Forest) confirmed that **alcohol is a key predictor in red wine and very important in white**, but performance improves when it's **combined with other features**.

Limitations:

- **Size of datasets**
- Models are based only on **physicochemical data** – they do not include **sensory features** (e.g. colour)
- The **quality label** is subjective and based on human scoring.

Patrycja Szostak

Tools & Methods Used

Exploratory Data Analysis (EDA):
- Python, Pandas, Seaborn, Matplotlib
- Distribution plots
- Correlation matrices
- Scatter plots and boxplots

Machine Learning Algorithms:
- Decision Tree Classifier
- Random Forest Classifier
- Train-test split (scikit-learn)

Evaluation Metrics:
- Accuracy score
- Confusion matrix
- Feature importance scores

Preprocessing and balancing:
- Feature selection (top 4 features)
- Ablation testing (removing/adding features)
- SMOTE (Synthetic Minority Oversampling Technique) for class imbalance

Questions