# Machine Learning for Wine Quality Classification

Comprehensive Evaluation & Analytical Insights

AUTHOR
Eshwaree Mathanki

# TABLE OF CONTENTS

## CONTEXT & SETUP


OVERVIEW 01
DATASET & PREPROCSSING 02
EXPLORATORY DATA ANALYSIS 03


## MODELING & EVALUATION


MODELING STRATEGY 04
RESULTS & PERFORMANCE ANALYSIS 05
VALIDATION & SIGNIFICANCE 06
RFEATURE IMPORTANCE. 07


## INSIGHTS & CLOSING


LIMITATIONS & RECOMMENDATIONS 08
SUMMARY 09
REFERENCES 10


01

Eshwaree Mathanki

# Overview

At-a-glance summary of methodology, results, and strategic insights

## INTRODUCTION

Predictive modeling of wine quality using physicochemical features (UCI Wine Quality dataset) to deliver interpretable, data-driven signals for stakeholder decision-making.

## OBJECTIVE

Build and evaluate ML models **to classify wine quality** into Low, Medium, and High categories using physicochemical features. Deliver robust, interpretable recommendations suitable for analytics-driven product and process improvement.

## PROCESS

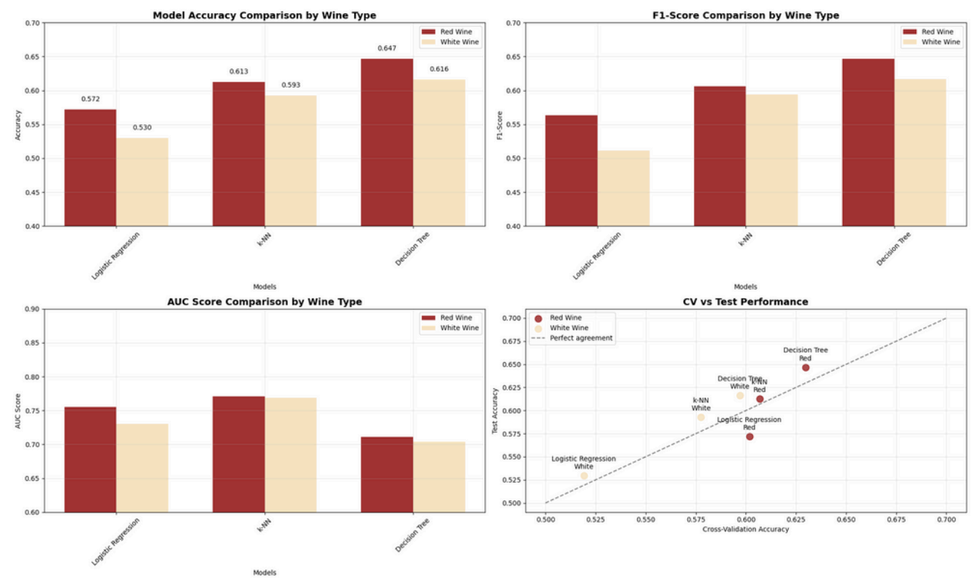
Combined binning, stratified splits, tuned k-NN & Decision Trees, 5-fold CV + ROC AUC, feature interpretability, and human-in-loop safeguards; optimized for internal quality analytics.

## SCOPE

Due to subjective labels and missing context, models serve for internal analytics and feature exploration, but not for consumer-facing automation or mission-critical QC without domain validation, calibration, and safeguards.

## KEY TAKEAWAYS

The framework achieved 64.7% accuracy for red wine and 61.6% for white wine classification, illustrating how ML approaches can extract patterns from physicochemical data. The practice nature of the dataset underscores the importance of domain validation for production applications.




## STRATEGIC IMPLICATION

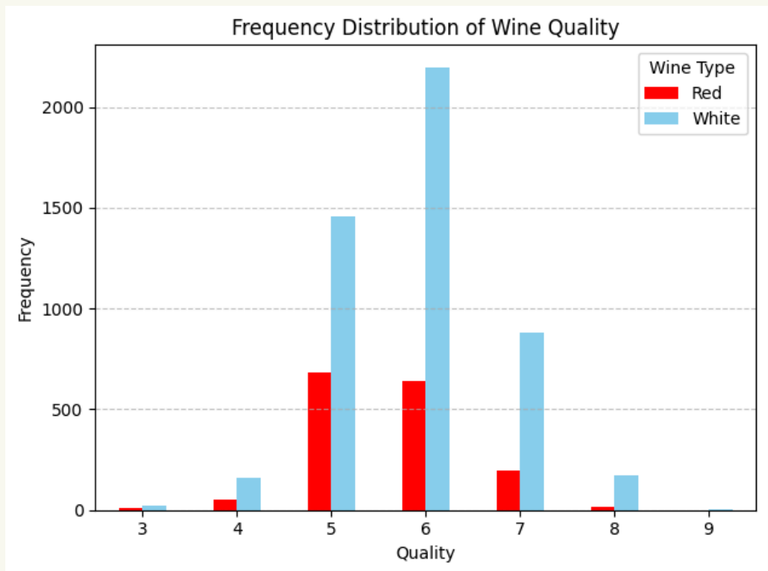
This project illustrates how explainable machine learning can enhance decision-making across the value chain. For producers and quality control teams, it enables smarter prioritization of lab tests and early detection of anomalous batches. For product managers, it informs assortment and pricing decisions, strengthening market positioning. By emphasizing interpretability and human-in-the-loop oversight, the model delivers risk-aware automation that balances efficiency with trust, making it fit for real-world deployment.

# Dataset & Preprocessing

## 2.1 DATASET SUMMARY

Features: 11 physicochemical measurements
Sample Sizes:

- Red wine: 1,599 samples × 11 features
- White wine: 4,898 samples × 11 features
- Data Quality: No missing values; clean, structured format




## 2.2 STRATEGIC BINNING: METHOD & RATIONALE

1. Why Combined Binning?
   a. Ensures "Low," "Medium," and "High" **quality correspond to identical numerical range**s across both wine types
   b. **Prevents inconsistent semantic meaning** when comparing models between datasets, enabling a fair, apples-to-apples performance comparison
2. Medium Bin Challenge
   a. Human scores cluster around middle values (5-6)
   b. **"Medium" represents a human-convenience label rather than chemically distinct category**
   c. This inherent ambiguity explains systematic performance drops in Medium classification
3. Robust Implementation
   a. safe_qcut function handles edge cases (duplicate scores, boundary conditions)
   b. Prevents binning failures while maintaining equal-frequency categories
   c. Prepared error handling for real-world data variability

## 2.3 VALIDATION FRAMEWORK & PREPROCESSING

Data Splitting**:** Stratified 60% train / 20% validation / 20% test

- Preserves class distribution across splits
- Red wine: Train 959, Validation 320, Test 320 samples

Feature Scaling: MinMaxScaler applied for consistent feature ranges

## TAKEAWAY

The "Medium Problem" Isn't a Model Failure – it's a reflection of real-world quality assessment complexity. Thus confirming our approach mirrors domain reality.

# Exploratory Data Analysis: Key Insights

## 3.1 QUALITY DISTRIBUTION PATTERNS

1. Central Tendency: Pronounced clustering around scores 5-6
2. Human Subjectivity: Reflects natural rating bias toward "average" scores
3. Modeling Impact: Medium category inherently challenging due to ambiguous boundaries

## 3.2 FOCUSED APPROACH

We **did not rely on a raw Pearson correlation matrix** because:
The target is ordinal and binned (human-scored quality → quantiles). Pearson correlations assume continuous linear relationships and may be misleading.
Many feature-quality relationships are non-linear or monotonic, and target labels are coarse.

## 3.3 FEATURE-QUALITY RELATIONSHIPS

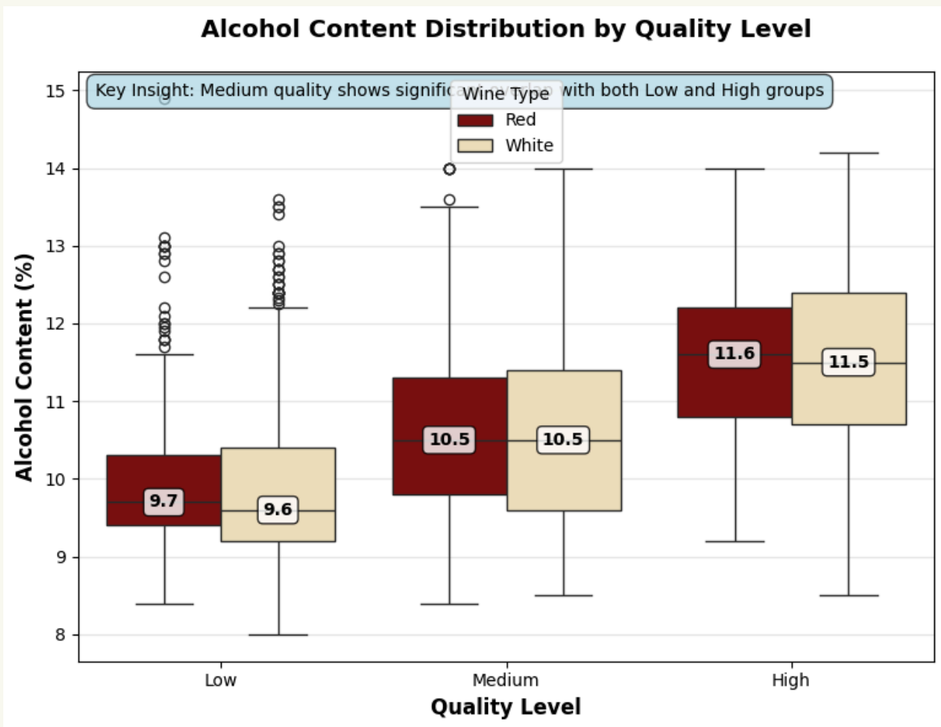
Key Pattern: Alcohol content shows strongest quality correlation but with significant Medium class overlap

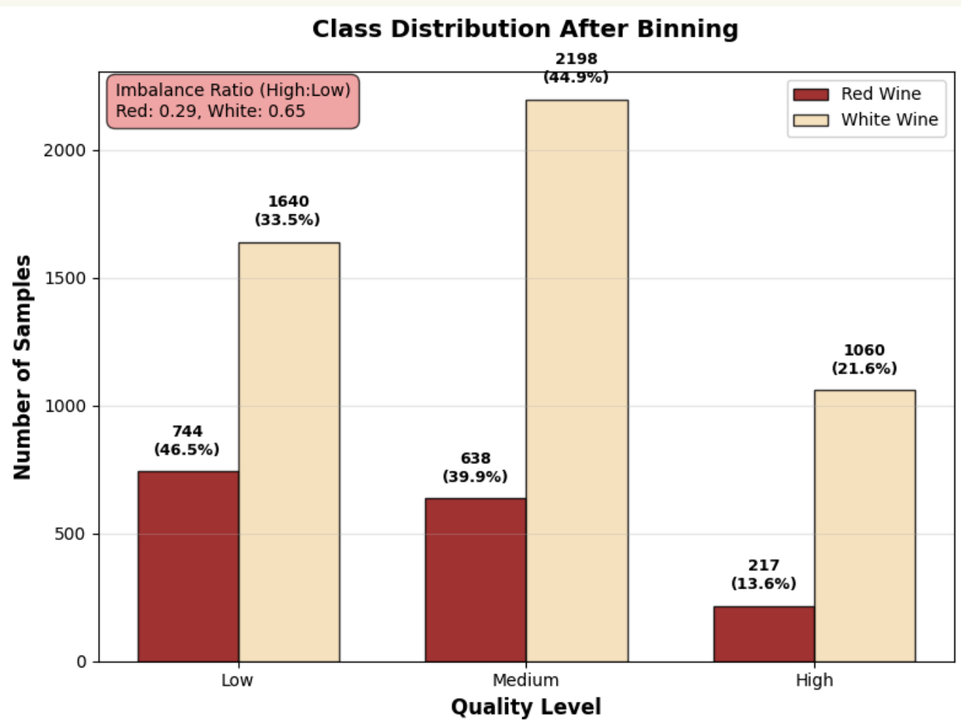
Note: This overlap highlights the difficulty of predicting “Medium”, it’s not a clear physical class but a human convenience label. Premium quality has identifiable chemical signatures, but Medium represents a ambiguous transition zone

## 3.4 CLASS BALANCE AFTER BINNING

Class Balance After Binning
- Red Wines: Low = 744, Medium = 638, High = 217
- White Wines: Low = 1640, Medium = 2198, High = 1060
- Implication: Strong class imbalance, particularly for red wines (fewer high-quality examples).

Thus, EDA confirms the strategic need for robust, explainable models that can handle fuzzy decision boundaries.



# Models: Methods & Strategy

## Logistic Regression

- Multinomial logistic regression with `class_weight='balanced'` to handle label imbalance
- 5-fold stratified CV for validation; baseline for testing linear separability
- Chosen for speed and interpretability; explains linear effects well, but struggles with non-linear overlap (esp. Medium class)

01.

## k-NN

- Distance-based classifier with tuned `k` for bias–variance tradeoff
- Scaling via MinMax ensures fair distance computation
- Useful for capturing local patterns; performs well on clustered classes, but sensitive to noise and class overlap

02.

## Decision Trees

- Depth-tuned, class-weighted trees to balance fit and generalization
- Inherently interpretable: feature splits highlight key drivers (e.g., alcohol)
- Strongest performer; captures non-linear relationships, but prone to overfitting without pruning

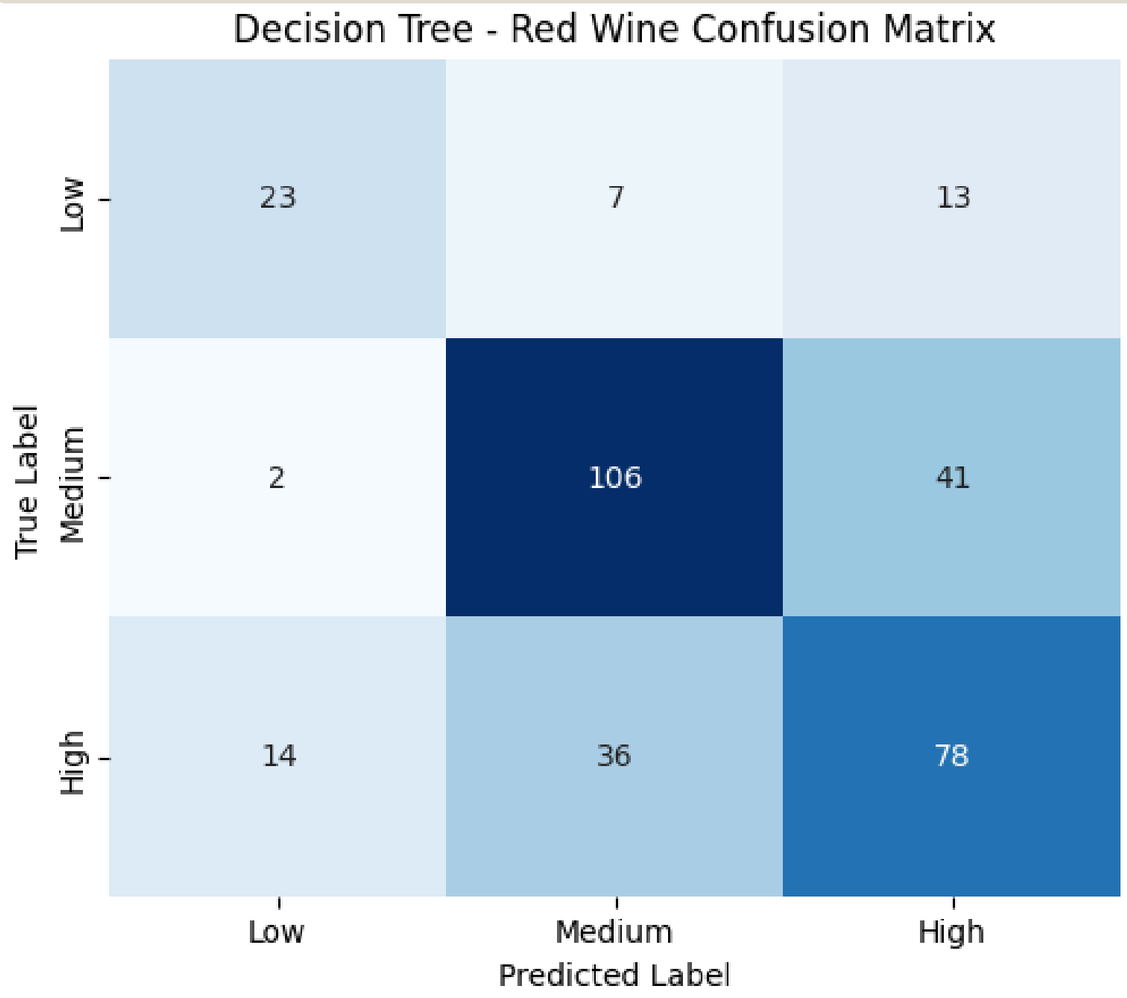
03.

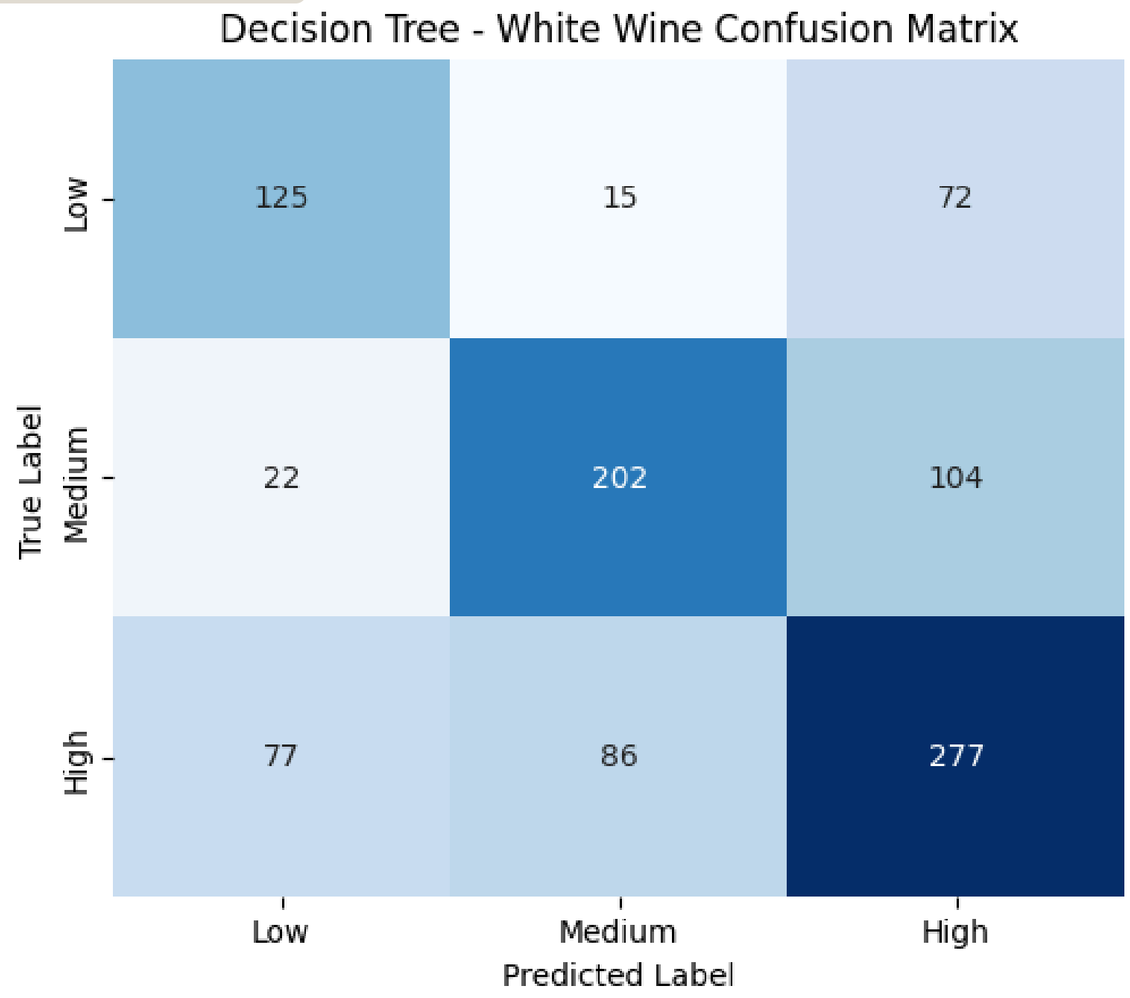
05

Eshwaree
Mathanki

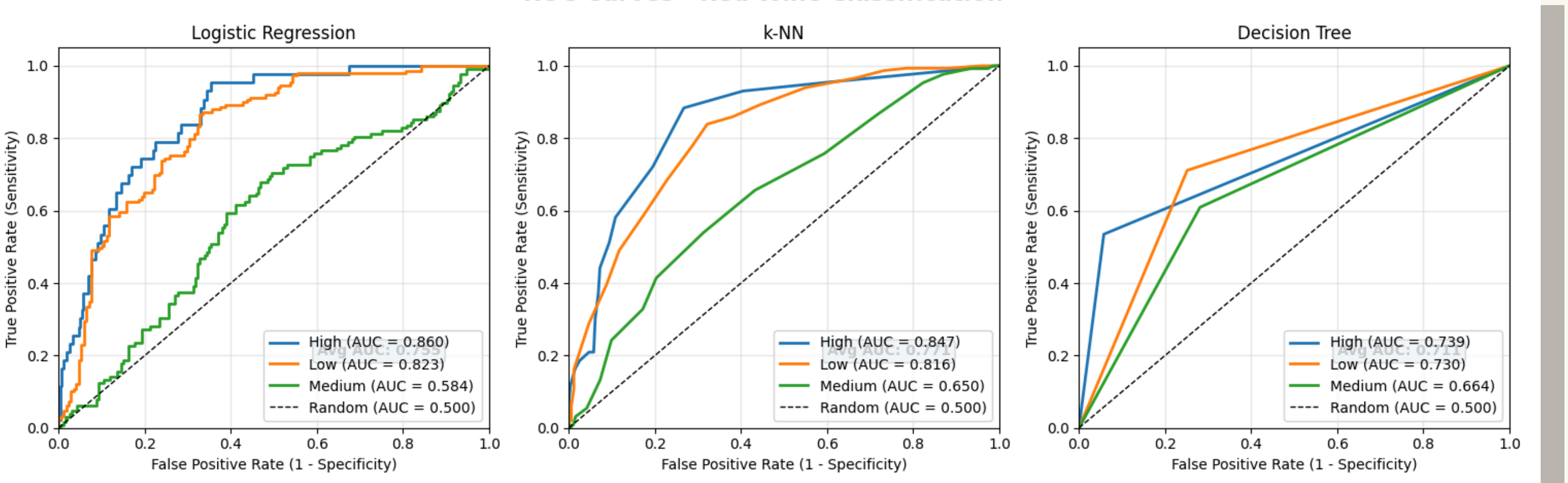
# Results & Analysis

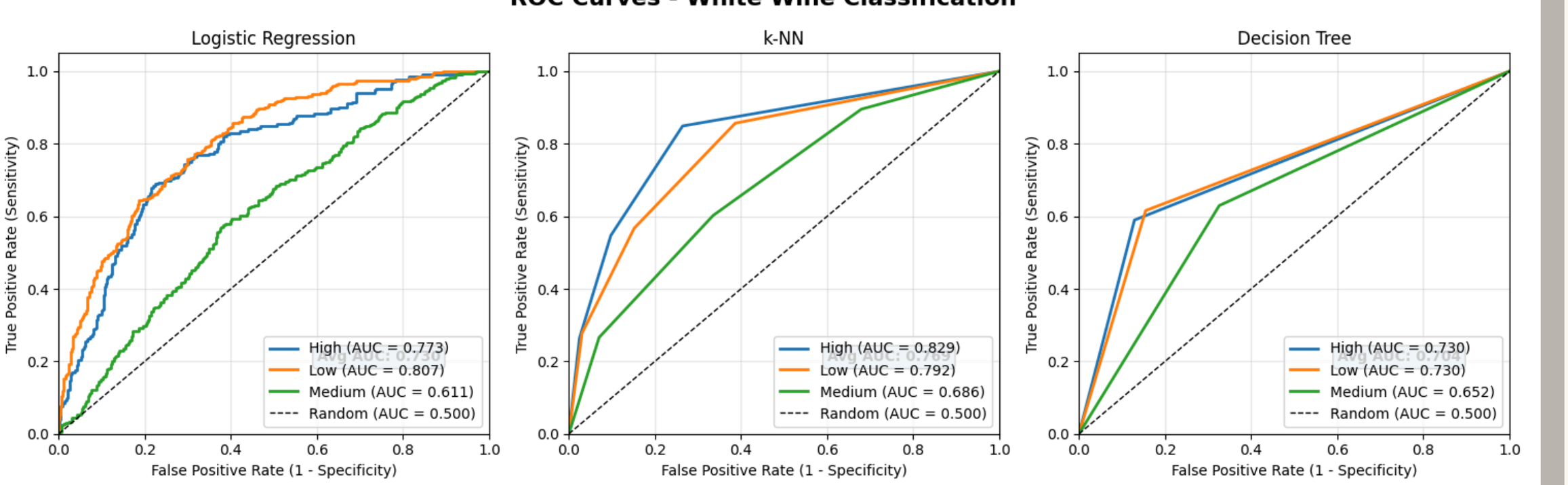
## 5.1 EVALUATION METRICES




Medium-quality class (middle row/column) shows the highest off-diagonal counts, confirming the class-overlap problem.



# Results & Analysis (contd.)

| Wine Type | Algorithm | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|---|
| Red | Logistic Regression | 0.5719 | 0.5832 | 0.5719 | 0.5635 |
| Red | k-NN | 0.6125 | 0.6025 | 0.6125 | 0.6061 |
| Red | Decision Tree | 0.6469 | 0.6469 | 0.6469 | 0.6466 |
| White | Logistic Regression | 0.5296 | 0.5492 | 0.5296 | 0.5117 |
| White | k-NN | 0.5929 | 0.5981 | 0.5929 | 0.5937 |
| White | Decision Tree | 0.6163 | 0.6184 | 0.6163 | 0.6169 |

k-NN offers the best probabilistic ranking ability (highest average AUC), while Decision Tree provides better discrete classification performance measured by F1.

## 5.2 CROSS-VALIDATION & STATISTICAL SIGNIFICANCE

5-fold Stratified Cross-Validation reduces overfitting risk and provides a more realistic estimate of model generalization

To test whether performance differences were meaningful rather than random:

- Paired t-tests were conducted on fold-level F1-scores
- Null hypothesis: no performance difference between models
- Significance level: $\alpha = 0.05$

Red Wine (Decision Tree vs. k-NN): t= 2.058, p = 0.1087 → Not significant
White Wine (Decision Tree vs. k-NN): t = 2.402, p = 0.0742 → Not significant

Interpretation: While Decision Trees outperform k-NN numerically, results are **not statistically significant** at the 95% confidence level.

| Wine Type | Algorithm | CV Score | Stability | Overall |
|---|---|---|---|---|
| Red | Logistic Regression | 0.6017 | 0.0158 | 0.612 |
| Red | k-NN | 0.6069 | 0.0163 | 0.6412 |
| Red | Decision Tree | 0.6298 | 0.029 | 0.6562 |
| White | Logistic Regression | 0.5191 | 0.0172 | 0.5296 |
| White | k-NN | 0.5776 | 0.0089 | 0.6253 |
| White | Decision Tree | 0.597 | 0.0162 | 0.6302 |

🎯 **Decision Trees remain the most practical choice** given their interpretability and robust performance, but statistical tests caution that observed improvements should be viewed as a directional signal rather than absolute superiority.

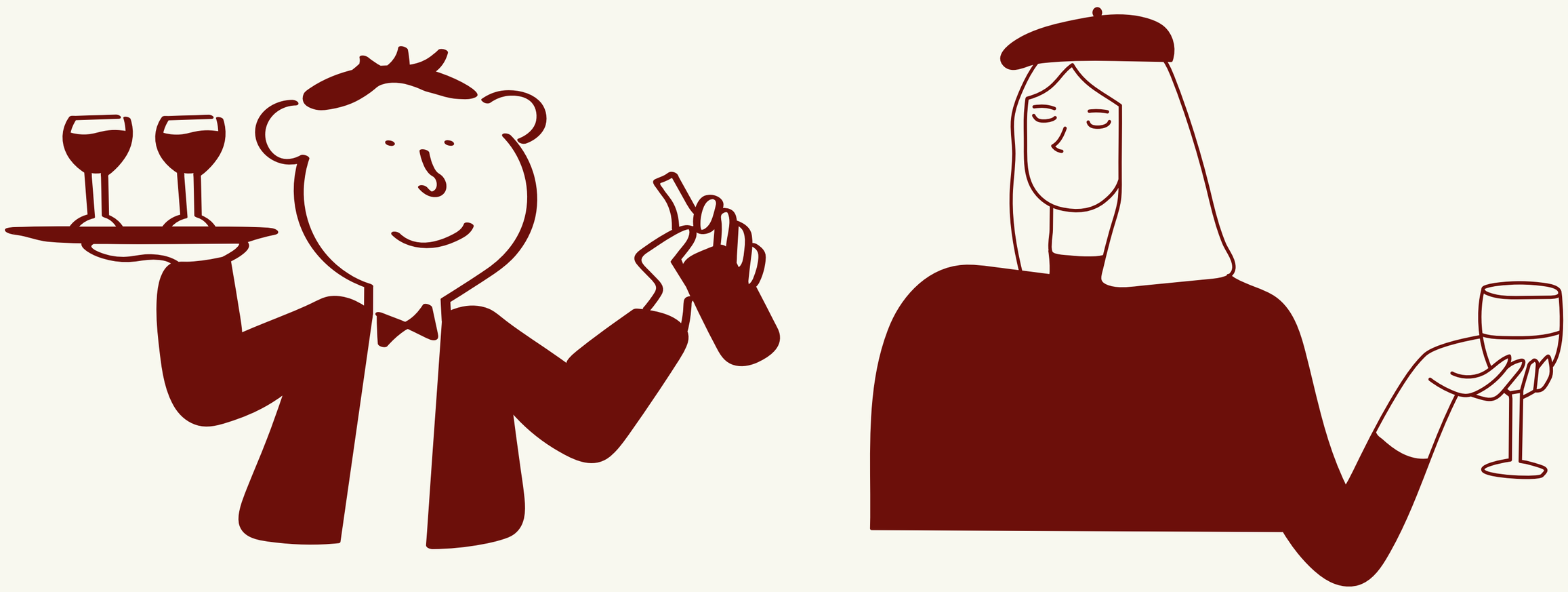
# Feature Importance

## 7.1 TOP 5 DECISION TREE FEATURE IMPORTANCES

| Feature | Red Wine |
|---|---|
| Alcohol | 0.2228 |
| Volatile acidity | 0.1405 |
| Sulphates | 0.0988 |
| Total sulfur dioxide | 0.0963 |
| Chlorides | 0.0819 |

| Feature | White Wine |
|---|---|
| Alcohol | 0.1477 |
| Volatile acidity | 0.1107 |
| Free sulfur dioxide | 0.1086 |
| Chlorides | 0.0952 |
| Total sulfur dioxide | 0.0933 |

Takeaway: Alcohol is consistently the single strongest predictor. These findings align with domain knowledge that alcohol often correlates with perceived body and richness, (Laguna et al., 2019).



# Limitations & Next Steps

## 8.1 DATA LIMITATIONS

- Inherent Subjectivity: Wine quality ratings reflect human judgment, which introduces bias and noise.
- Class Imbalance: Medium-quality wines are over-represented, skewing learning and performance metrics.
- Boundary Compression: The “Medium” category is effectively a single score (6), creating a narrow, heterogeneous group.
- Unknown Collection Details: Metadata on scoring methods, panel composition, and calibration procedures is unavailable, limiting confidence in label consistency.
- Limited Scope: Dataset covers only certain wine types and physicochemical features, restricting generalizability to other wine varieties or production contexts.

## 8.2 MODEL LIMITATIONS

- Moderate Performance: Accuracy (60-65%) leaves room for improvement; models struggle most with the Medium class.
- Business Risk: Misclassification of premium wines could lead to flawed quality or pricing decisions if deployed without safeguards.
- Interpretability Trade-offs: Decision Tree models achieve the best balance of F1-Score and accuracy, but exhibit higher cross-validation variance, indicating potential overfitting.
- Non-Linearity Constraints: Logistic Regression is limited by linear assumptions; k-NN requires careful tuning and can be sensitive to feature scaling.

## 8.3 NEXT STEPS: DATA ENHANCEMENTS

- Richer Labels: Multi-rater sensory panels or standardized scoring to reduce subjectivity.
- Expanded Features: Include chemical or production metrics to improve model discrimination.
- Data Provenance: Document score calibration, collection methods, and sample metadata to strengthen reliability.

## 8.4 NEXT STEPS: MODELING IMPROVEMENTS

- Ensembles & Interpretability: Random Forest or Gradient Boosting with SHAP values for robust, explainable performance.
- Advanced Validation: Nested cross-validation for unbiased hyperparameter tuning and generalization assessment.
- Class-Aware Adjustments: SMOTE or stratified resampling to mitigate Medium-class compression.
- Feature Insights: Partial dependence plots and sensitivity analysis for non-linear interactions and business-critical predictions, along with feature engineering to extract clearer insight

## CONCLUDING INSIGHTS

Hence, these measures ensure ML outputs are actionable, interpretable, and aligned with business priorities, enabling informed decision-making while maintaining risk-aware oversight.

09

Eshwaree
Mathanki

# Summary


**1.**
The analysis reveals that wine quality, as assessed by human raters, is most strongly influenced by alcohol content, which consistently drives perceived richness and body across both red and white wines

**2.**
The greatest challenge lies in medium-quality wines: their chemical profiles overlap with both higher and lower tiers, making them difficult to separate and reflecting the real-world ambiguity of what consumers consider “average.”



**THUS,**
The subjectivity of human scoring and the narrow range of mid-tier wines remain fundamental barriers to precision. Together, the results underscore both the promise and the limits of predictive ML in supporting quality evaluation.

# REFERENCES

UCI Machine Learning Repository. Wine Quality Dataset. https://archive.ics.uci.edu/ml/datasets/wine+quality

Han, J., Pei, J., & Tong, H. Data Mining: Concepts and Techniques, 4th Edition. Morgan Kaufmann, 2020.

Laguna, L., Álvarez, M. D., Simone, E., Moreno-Arribas, M. V., & Bartolomé, B. (2019). Oral Wine Texture Perception and Its Correlation with Instrumental Texture Features of Wine-Saliva Mixtures. Foods, 8(6), 190. https://doi.org/10.3390/foods8060190

# Machine Learning for Wine Quality Classification

## Comprehensive Evaluation & Analytical Insights

AUTHOR
Eshwaree Mathanki