Binary Classification of White Wine Quality

Predicting wine quality using supervised machine learning

Masha S. Logan

Why Predict Wine Quality?

1. Traditional Wine Assessment is time-consuming and prone to human error.
2. Data-Driven Prediction enables faster, more efficient decision-making.
3. Cost Efficiency helps avoid wasting resources on subpar batches.
4. Timely Interventions allow winemakers to optimize production and prevent costly errors.
5. Anticipate market reception and optimize pricing strategies.

Can we classify white wine as high-quality or low-quality based on its chemical properties using supervised learning?

Data Preprocessing & Challenges

Dataset

- UCI Wine Quality Dataset
- 4,898 white wine samples
- 11 physicochemical properties (e.g., acidity, sugar, alcohol
- No missing values

Converted **wine quality scores (0-10)** into **binary labels**

- High-quality (1): Scores ≥7
- Low-quality (0): Scores ≤6

Class imbalance issue →

- Only **22%** of wines are high quality

Key Challenge:

Models may struggle to correctly classify **high-quality wines** due to **imbalance**

Exploratory Data Analysis

Strong correlations:

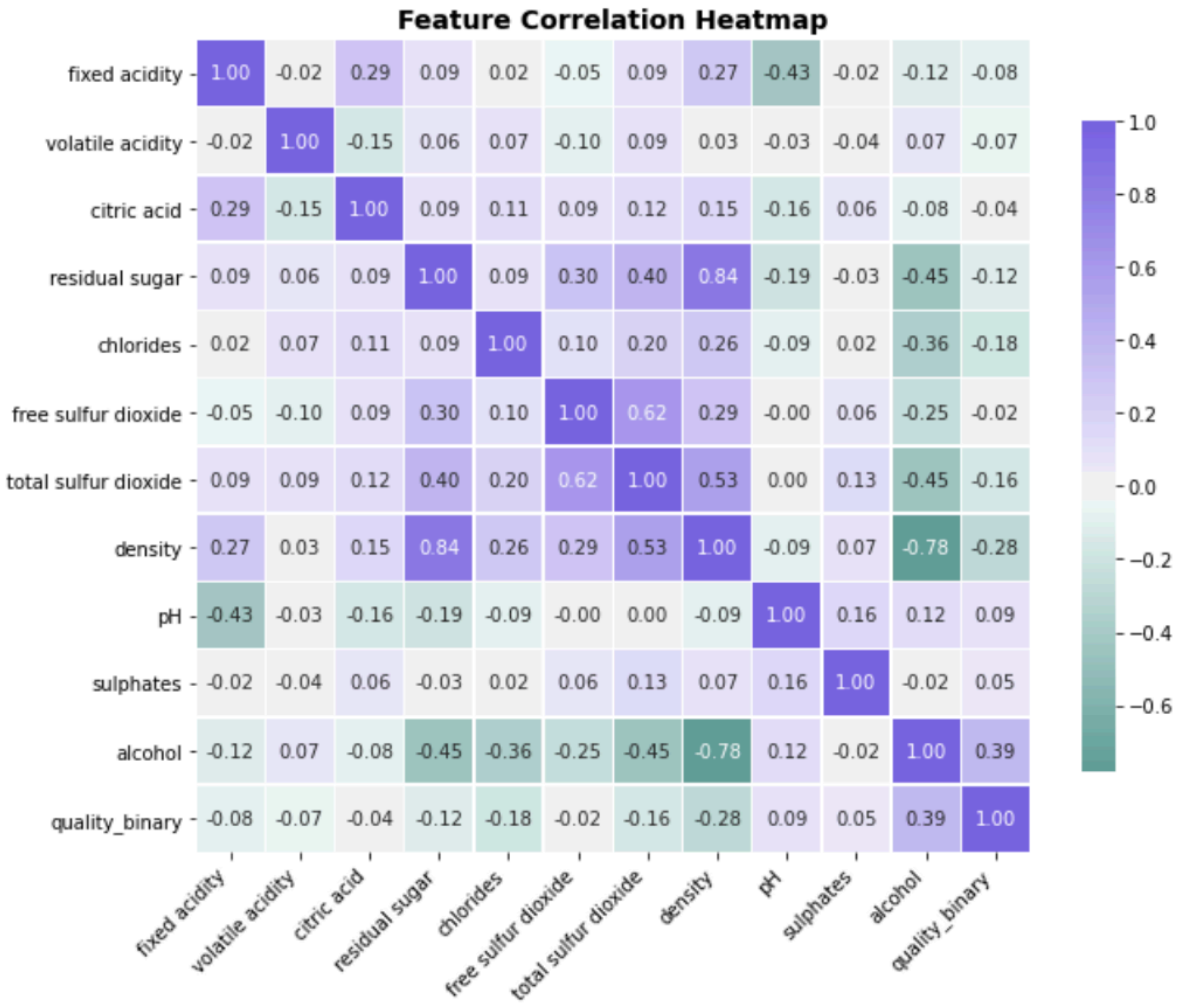
- density and residual sugar
- alcohol content and density

Moderate correlations:

- sulfur dioxide compounds

High quality wines tend to have:

- higher average alcohol content (11.42% vs 10.26%)
- lower levels of total sulfur dioxide
- slightly lower residual sugar levels



Machine Learning Models

Supervised Learning Models

1. Logistic Regression
2. Random Forest
3. Support Vector Machine

Model Evaluation

- Accuracy
- Precision
- Recall
- F1-Score

Additional details

- Using RobustScaler for feature scaling
- Stratified K-Fold Cross-Validation for reliable evaluation
- Hyperparameter tuning using RandomizedSearchCV

Machine Learning Models

High-quality wine classification remains a challenge

Logistic Regression performs the weakest with **71.2% accuracy**

- **decent recall for high-quality wines (76.0%)**
- poor **precision (41.0%,)** resulting in many false positives

Random Forest is the best performer with **88.1% accuracy**

- it balances **precision (81.7%)** and **recall (58.2%)** for high-quality wines
- still misclassifies **41.8%** of them as low-quality

SVM shows **high accuracy (86.1%)**

- favors low-quality wines with **99.9% recall**
- only **36.04%** of high-quality wines are correctly classified.

Model	Accuracy	Low Quality Wines			High Quality Wines		
		Precision	Recall	F1	Precision	Recall	F1
Logistic Regression	71%	91%	70%	79%	41%	76%	53%
Random Forest	88%	89%	96%	93%	82%	58%	68%
SVM	86%	85%	100%	92%	99%	36%	53%

Adressing Class Imbalance with SMOTE

High-quality wine classification remains a challenge

Synthetic Minority Over-sampling Technique (SMOTE)

- generates **synthetic samples** for the **minority class** instead of duplicating existing data.
- creates a **more balanced dataset**, enabling models to better learn patterns

Impact of SMOTE

- Improves recall for high-quality wines, reducing the bias toward low-quality wines
- Ensures that models perform more effectively in real-world scenarios where class imbalance is common

Model Performance after SMOTE

Handling class imbalance effectively is crucial for achieving balanced model performance

Logistic Regression (SMOTE) shows minor improvements

- Recall for high-quality wines went up from 75.9% to 77.2%, but accuracy remains lower at 73.8%.

Random Forest (SMOTE) is the best overall performer

- 92.4% accuracy and 94.6% recall for high-quality wines, and balances precision precision and recall well

SVM (SMOTE) struggled

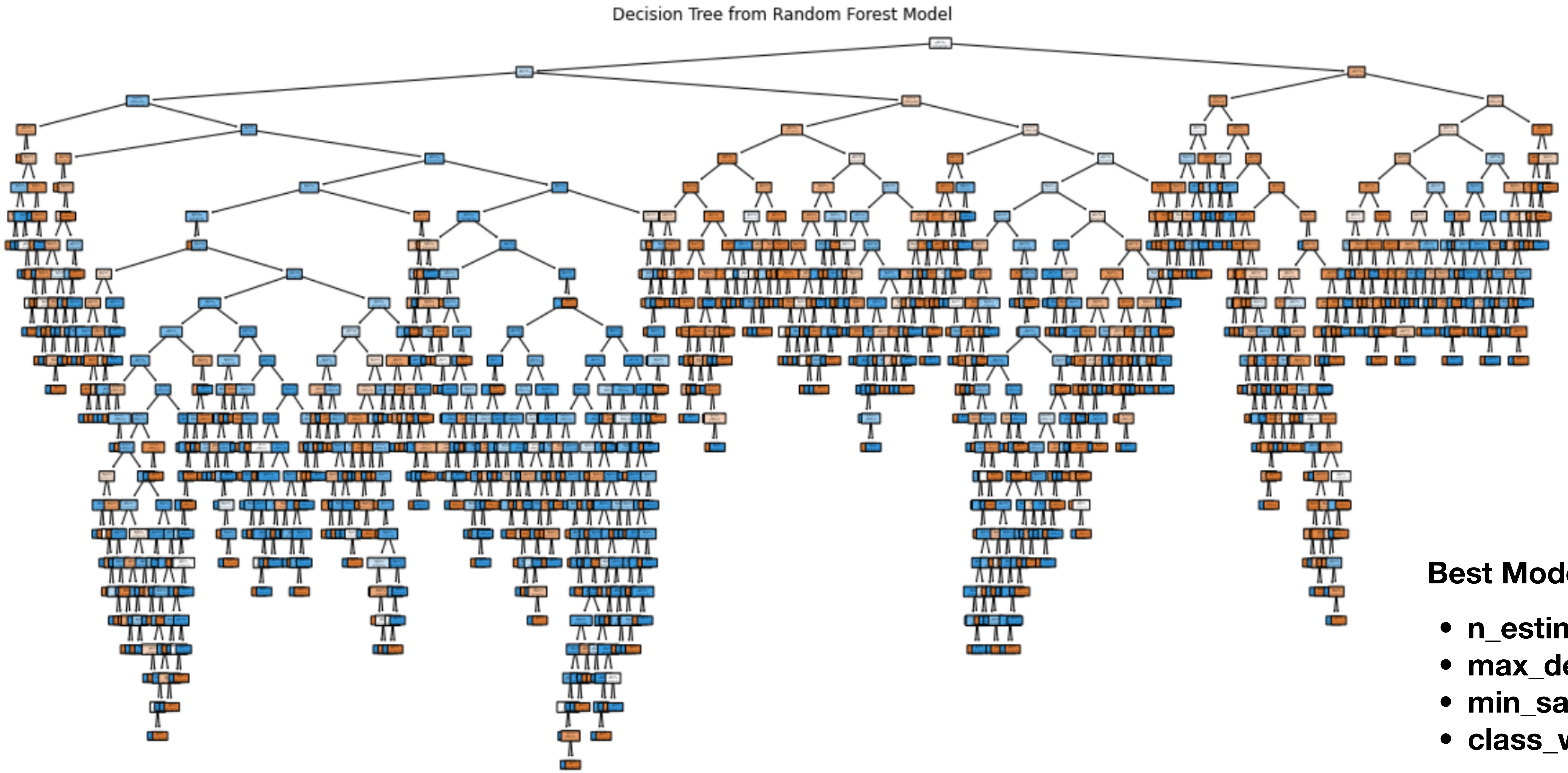
- recall for high-quality wines dropped to 11.6%, with accuracy dropping to 55.8%

Model	Accuracy	Low Quality Wines			High Quality Wines		
		Precision	Recall	F1	Precision	Recall	F1
Logistic Regression	71%	91%	70%	79%	41%	76%	53%
Logistic Regression (SMOTE)	74%	76%	70%	73%	72%	77%	75%
Random Forest	88%	89%	96%	93%	82%	58%	68%
Random Forest (SMOTE)	**92%**	**94%**	90%	**92%**	**91%**	**95%**	**93%**
SVM	86%	85%	100%	92%	99%	36%	53%
SVM (SMOTE)	56%	53%	100%	69%	100%	12%	21%

SMOTE can significantly improve performance in addressing class imbalance.

Random Forest (SMOTE) - Decision Tree Visualization

Just one tree from the Forest



Best Model Parameters

- **n_estimators**: 300
- **max_depth**: None
- **min_samples_split**: 2
- **class_weight**: 'balanced'

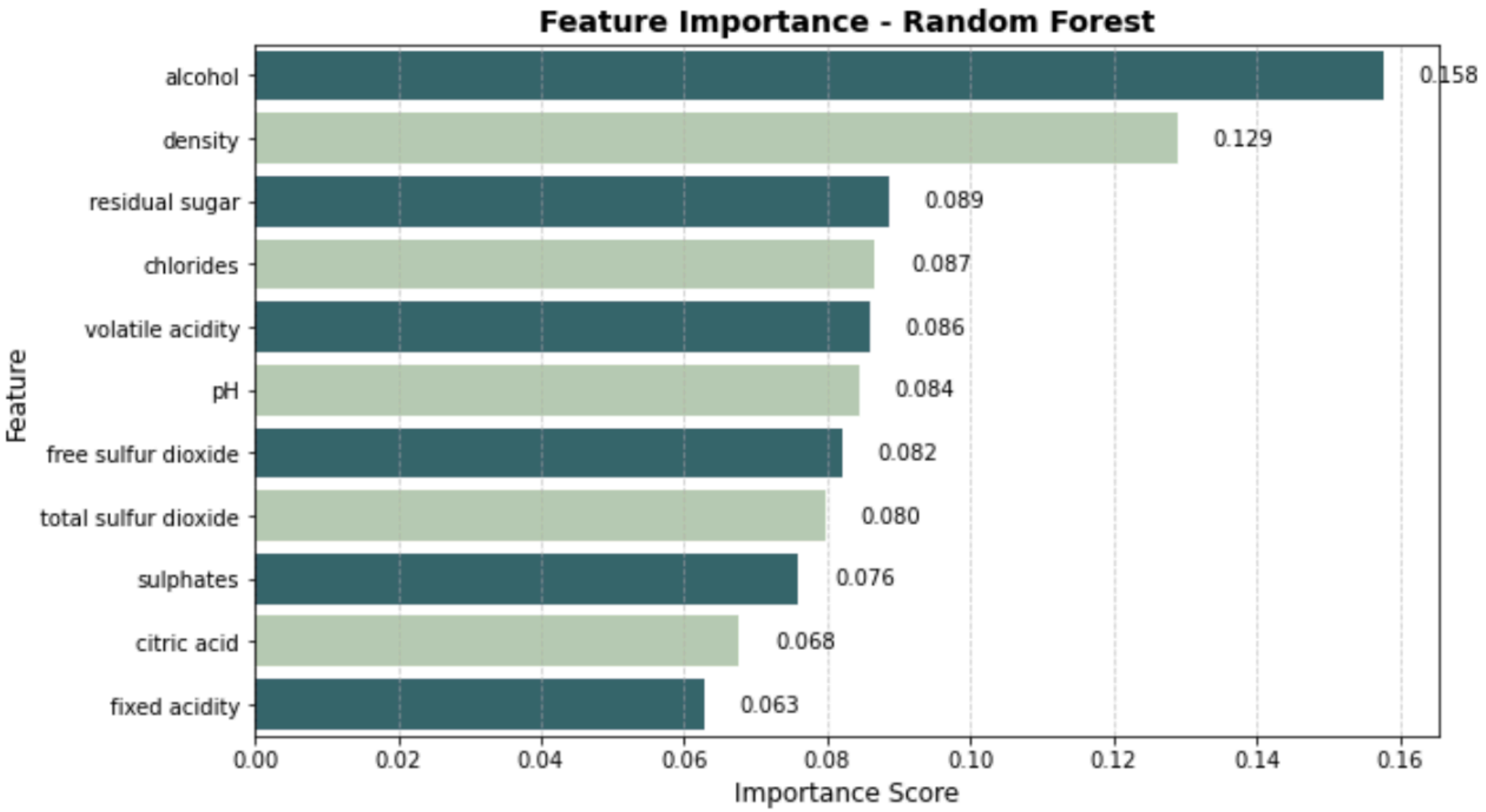
Random Forest (SMOTE) - Feature Importance

Alcohol content:
- most influential predictor of wine quality
- higher alcohol wines rated as higher quality

Density
- likely influenced by interactions between **sugar content** and **alcohol levels**

Other significant predictors:
- **Residual sugar a**ffects sweetness and balance
- **Chlorides** contribute to the stability and perception of wine
- **Volatile acidity** impacts the balance of acidity in wine.

Fixed acidity and **citric acid**:
- minimal influence, indicating they are less discriminative for wine quality.



Conclusion and Future Work

Random Forest with SMOTE provides a **reliable framework** for classifying white wine quality based on physicochemical attributes.

- offers valuable insights for winemakers to predict quality during production
- enables **proactive quality management**

Most influential predictors of wine quality:

- Alcohol content and density
- Residual sugar, volatile acidity, and chlorides

Future Work:

- Explore models like **XGBoost** or **LightGBM** for better performance.
- **Feature engineering** (e.g., alcohol-to-sugar ratio)
- Experiment with **feature engineering** (e.g., alcohol-to-sugar ratio) and alternative **class balancing techniques** like **SMOTE-Tomek**.

Thank you!

Please reach out with any questions and suggestions:

LinkedIn: https://www.linkedin.com/in/mashalogan/
GitHub: https://github.com/mashuzza

Project files: https://github.com/mashuzza/python-projects/tree/main/supervised-learning-wine-quality-classification