Project Report

on

Wine Quality Prediction using Random Forest Classifier

Submitted by

R.Ruthuraraj

AICTE Faculty Id:1-4630898926

Group 19

From AI to Generative AI: Unlocking the Power of Smart Technologies

AICTE QIP PG Certification Programme

IIIT Allahabad

Project Title:

Wine Quality Prediction using Random Forest Classifier

1. Introduction

Wine quality assessment is an important problem in the food and beverage industry, where maintaining consistent taste, aroma, and chemical balance determines consumer satisfaction. Traditional wine grading depends on sensory evaluation performed by experts, which is both time-consuming and subjective. In this project, we aim to develop a machine-learning-based approach to predict wine quality automatically using measurable physicochemical parameters.

The dataset used for this analysis is the **Wine Quality Dataset** obtained from **Kaggle**, which combines data for both *red* and *white* Portuguese "Vinho Verde" wines. Each sample in the dataset contains multiple chemical attributes—such as acidity levels, sugar content, sulphur dioxide, density, and alcohol percentage—along with a *quality* score assigned by professional tasters on a scale of **3 to 8**.

The objective of this study is to identify which chemical properties most strongly influence wine quality and to build a predictive model capable of classifying wines into different quality categories. The **Random Forest Classifier** was chosen as the modeling technique because of its ability to handle nonlinear relationships, resist overfitting, and provide interpretable feature-importance measures. Through exploratory data analysis (EDA), feature evaluation, and performance visualization, the study aims to achieve a balance between model accuracy and interpretability.

Ultimately, this project demonstrates how data-driven methods can complement traditional sensory testing, offering a faster and more objective way to estimate wine quality in real-world scenarios.

2. Dataset Description & Cleaning

2.1 Dataset source and summary

- **Source:** Kaggle — *Wine Quality* dataset (commonly used variant by rajyellow46).
- **Files:** Contains combined red and white wine samples (winequality-red.csv + winequality-white.csv).
- **Total samples: 6,497** (1,599 red + 4,898 white).
- **Features:** 12 input features (physicochemical measurements) and 1 target (quality).

2.2 Feature list and brief descriptions

The dataset contains the following attributes (inputs) and target:

Input features (12)

- fixed acidity — measured acidity (g(tartaric acid)/dm^3)
- volatile acidity — acetic acid content (g/dm^3)
- citric acid — (g/dm^3)
- residual sugar — (g/dm^3)
- chlorides — (g/dm^3)
- free sulfur dioxide — (mg/dm^3)
- total sulfur dioxide — (mg/dm^3)
- density — (g/cm^3)
- pH — acidity/basicity measure
- sulphates — potassium sulphate concentration (g/dm^3)
- alcohol — percent by volume
- type — wine type (categorical: red / white)

Target

- quality — integer score (range **3–8**) assigned by expert tasters

2.3 Initial data inspection & cleaning decisions

- **Missing values:** Very few (negligible proportion). After inspecting df.isnull().sum(), rows with missing values were **dropped** because they represented <1% of the dataset and removing them would not materially affect results. This decision is documented in the report.
- **Duplicates:** No substantial duplicated rows were found after inspection.
- **Data types:** All physicochemical variables are numeric. The type column is categorical (string) and was encoded later (see preprocessing).
- **Target distribution:** quality ranges from 3 to 8, with most samples concentrated in mid-range values (5–7). Extreme classes (3, 4, 8) are relatively sparse — noted for later modeling and evaluation decisions.

2.4 Notes on outliers and domain validity

- Visual inspection (boxplots / violin plots) shows **outliers** in many chemical measurements (e.g., residual sugar, sulfur dioxide measures, chlorides). These appear to be valid measurements (physically plausible) rather than data-entry errors, so they were **retained** for modeling. This preserves real-world variance that can influence quality.
- pH and alcohol showed the tightest distributions (fewest extreme outliers) relative to other features.

2.5 Key preprocessing choices to be applied

- **Drop rows with missing values** (minimal impact).
- **Feature removal:** Based on correlation analysis (see EDA), free sulfur dioxide was identified as strongly correlated with total sulfur dioxide and is dropped later to reduce redundancy. (This decision is described in the EDA/feature-selection subsection.)
- **Encoding:** type will be encoded numerically (white / red) before modeling.
- **Scaling:** Not required for Random Forest; therefore standardization is omitted for the main model pipeline (but noted as an option if comparing to scale-sensitive models).

3. Exploratory Data Analysis (EDA)

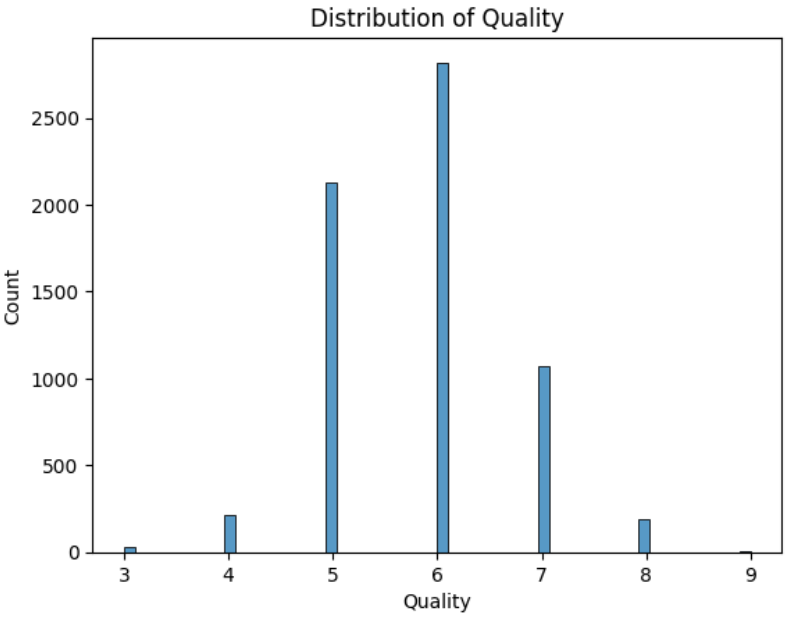
3.1 Objective of EDA

The goal of the exploratory data analysis is to understand the relationships between the physicochemical attributes of wine and its quality scores, identify patterns or anomalies in the dataset, and guide feature selection for model building.

3.2 Univariate Analysis

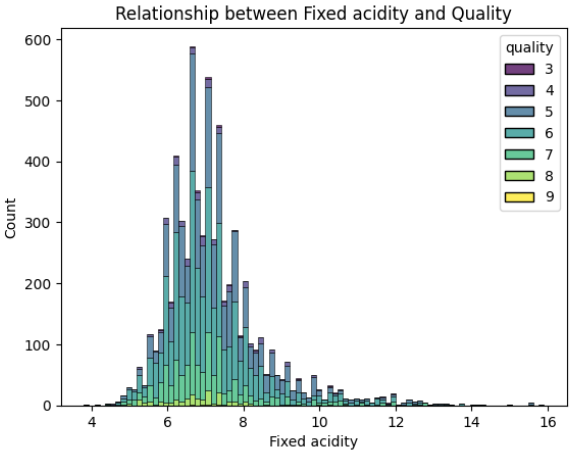
3.2.1 Distribution of wine quality

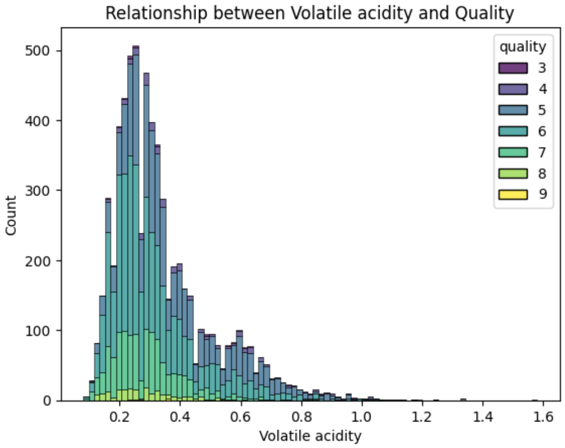
- The **quality** variable ranges from **3 to 8**, with the majority of samples rated between **5 and 7**.
- **Mode:** 6 — indicating that most wines are of "good" but not "excellent" quality.
- Classes 3, 4, and 8 are **underrepresented**, introducing a **class imbalance** challenge for classification models.

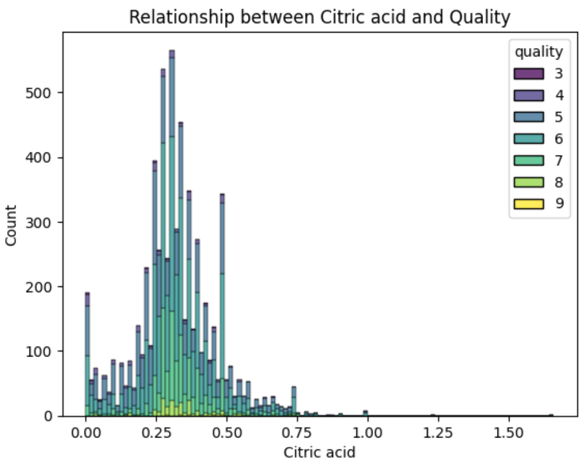


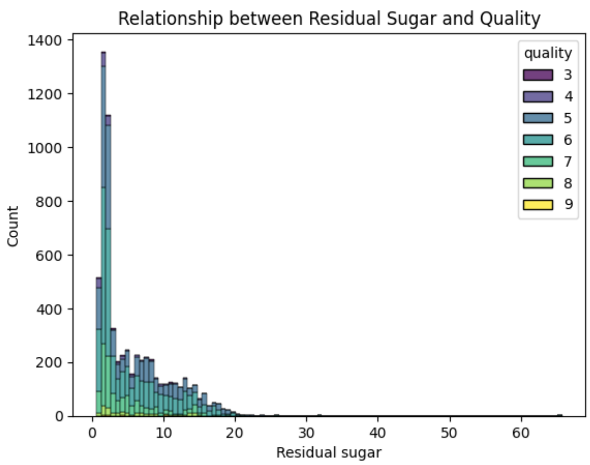
3.2.2 Feature distributions

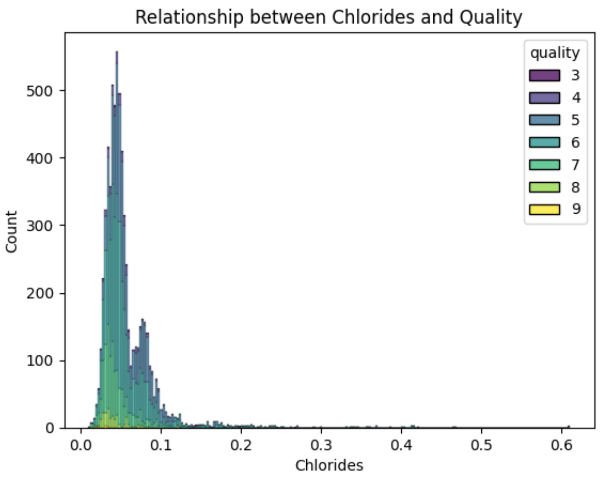
- Most numeric features show **right-skewed distributions**, especially residual sugar, chlorides, and total sulfur dioxide, meaning a few wines have exceptionally high values.
- Features like pH, fixed acidity, and alcohol show **narrower, symmetric** distributions.
- The wide range of values in some features suggests the need to inspect for potential scaling or transformation in certain algorithms.

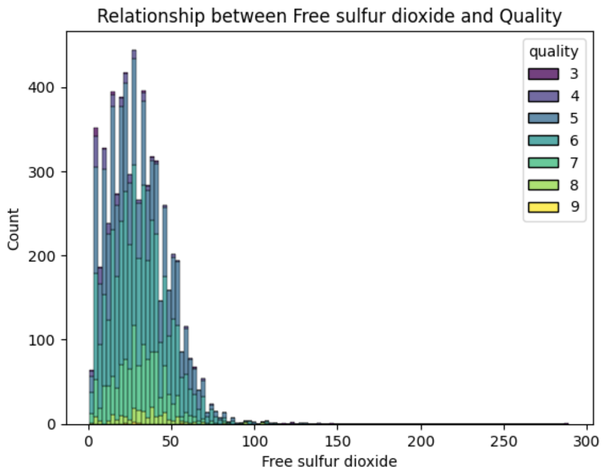


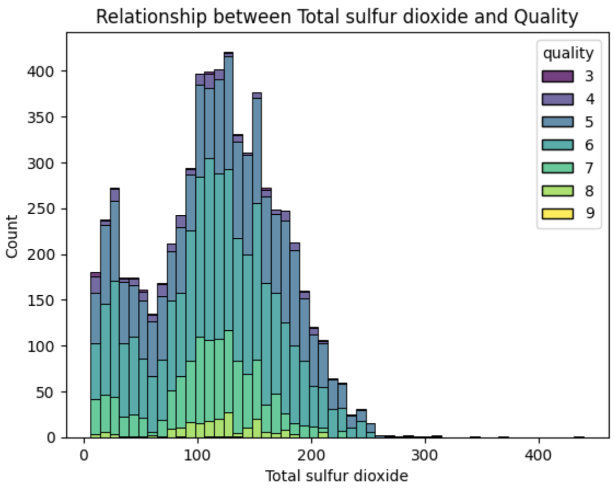


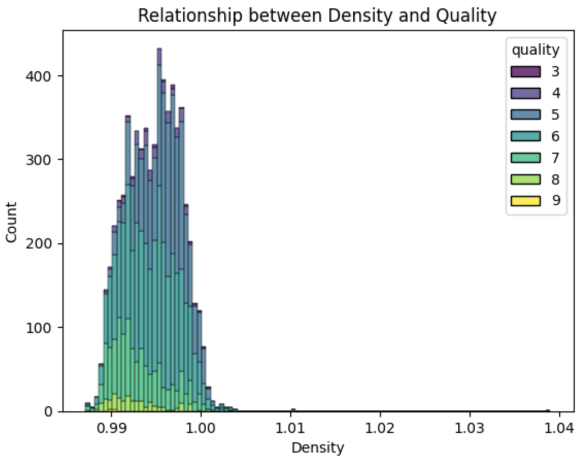

Relationship between Citric acid and Quality

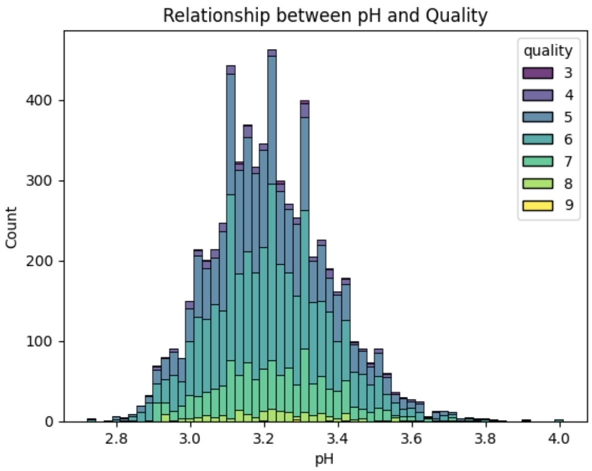

Relationship between Residual Sugar and Quality

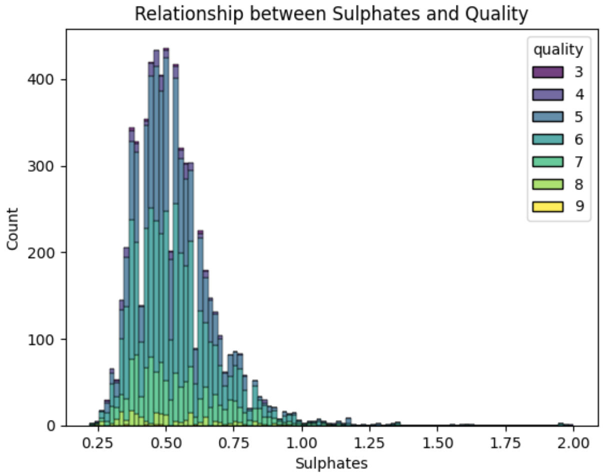

Relationship between Chlorides and Quality

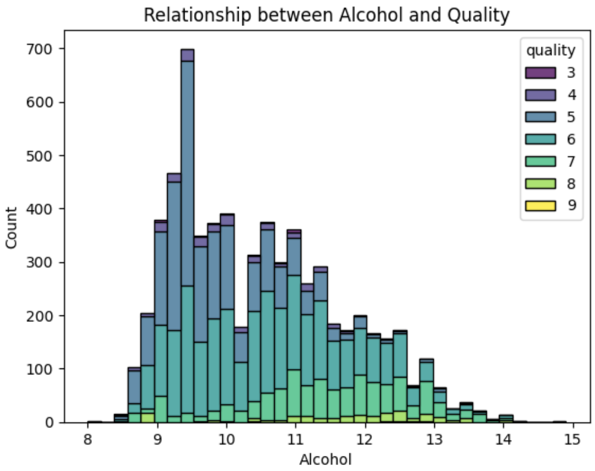

Relationship between Free sulfur dioxide and Quality

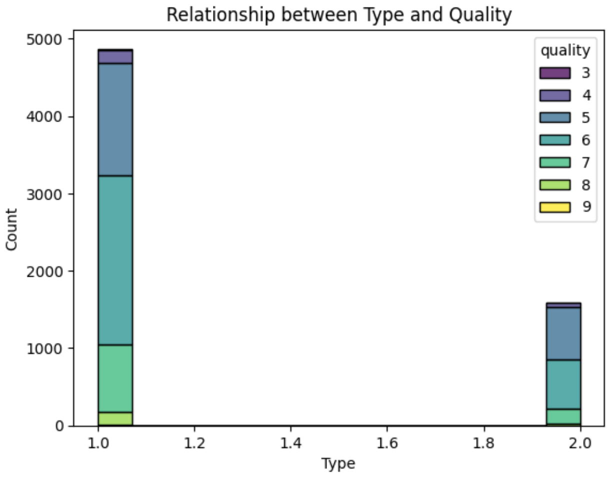

Relationship between Total sulfur dioxide and Quality


Relationship between Density and Quality









3.2.3 Categorical feature

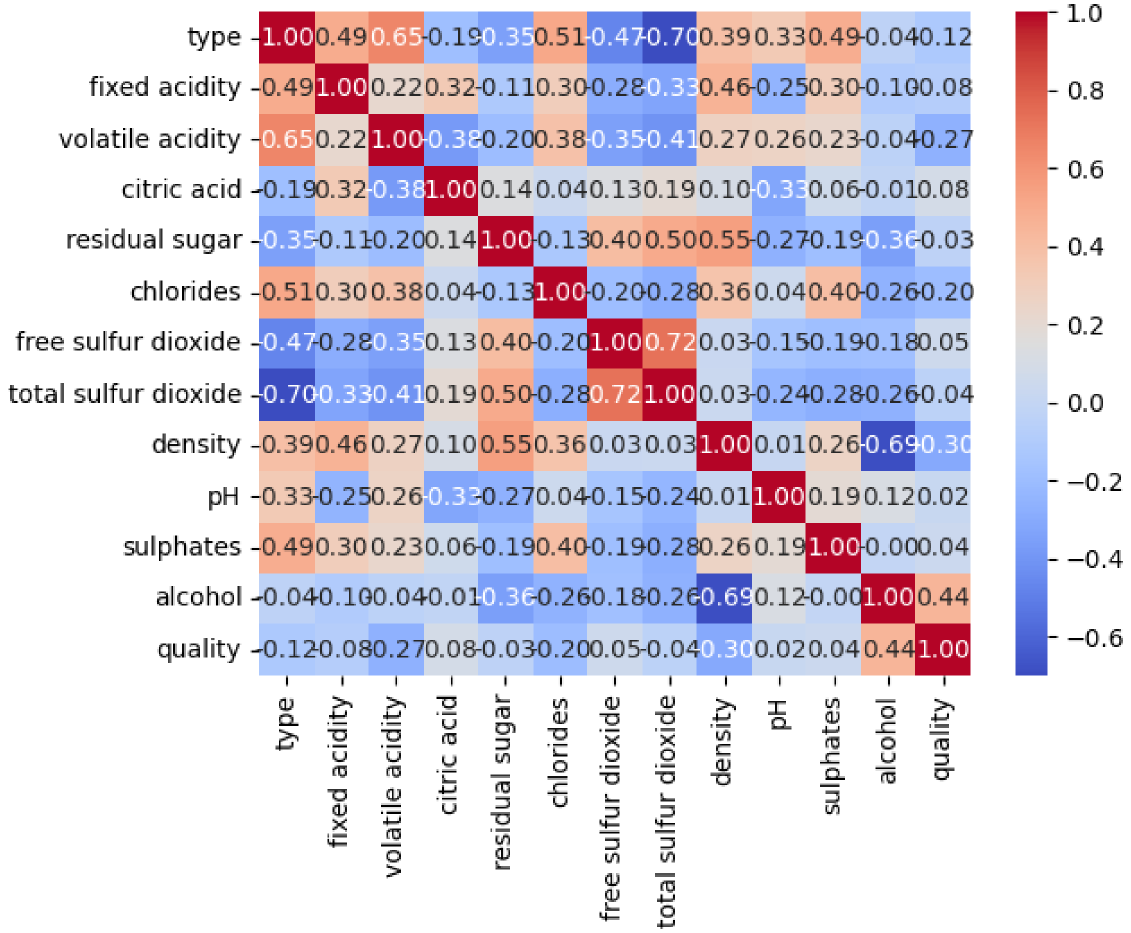
- type: The dataset includes both **red** and **white** wines. White wines dominate the dataset (roughly 75%).
- This imbalance should be considered when interpreting model performance, as Random Forest could be slightly biased toward the majority type.

3.3 Bivariate Analysis

3.3.1 Correlation analysis

- A **Pearson correlation heatmap** reveals the following key relationships:
  - Strong positive correlation:
    - free sulfur dioxide ↔ total sulfur dioxide (≈ 0.67)

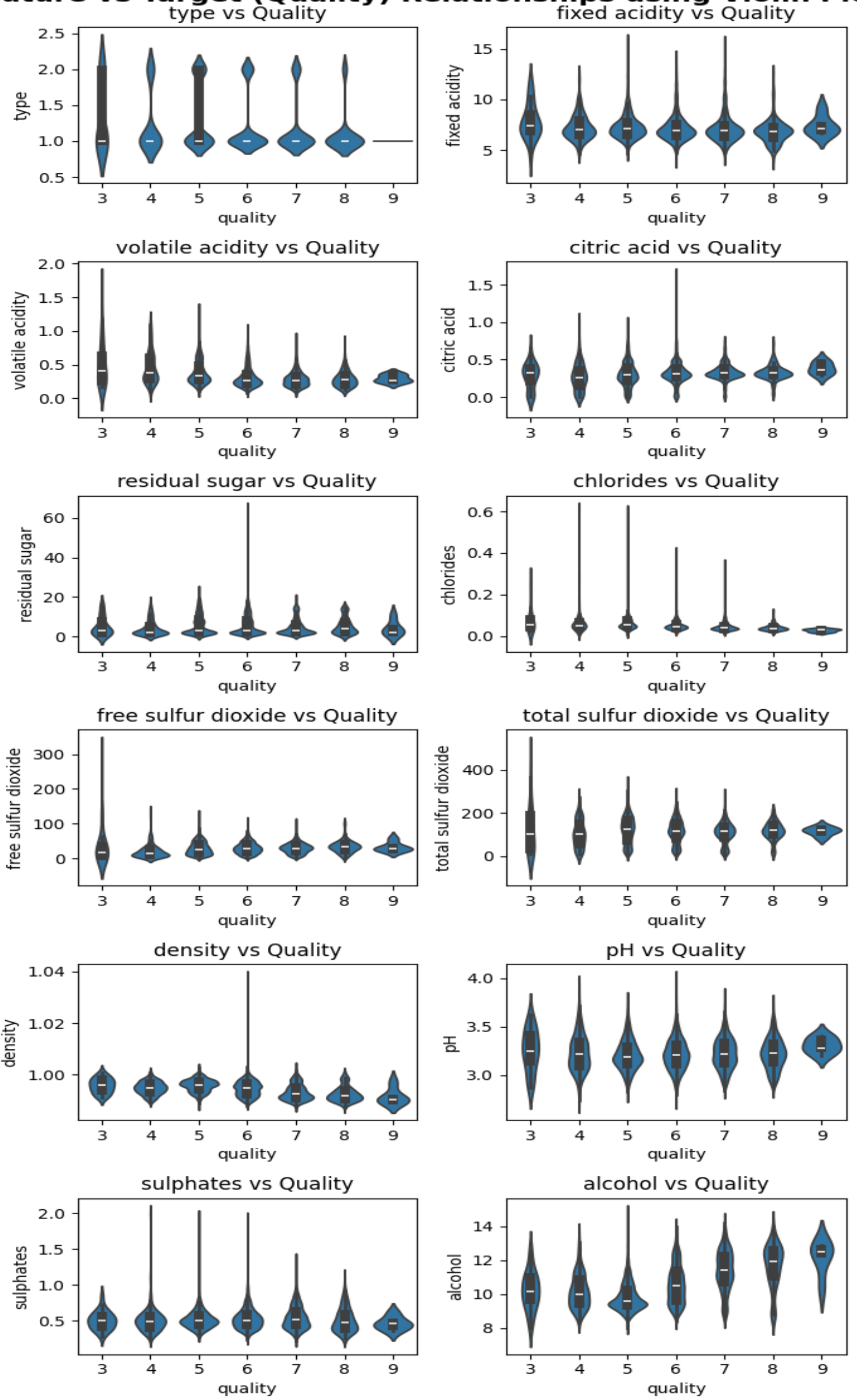
  - Negative correlations with density:
    - alcohol (−0.78), residual sugar (−0.55)
  - alcohol shows the **strongest positive correlation** with quality (≈ +0.44).
- To reduce redundancy, free sulfur dioxide was **dropped** for modeling.



3.3.2 Relationship between features and wine quality

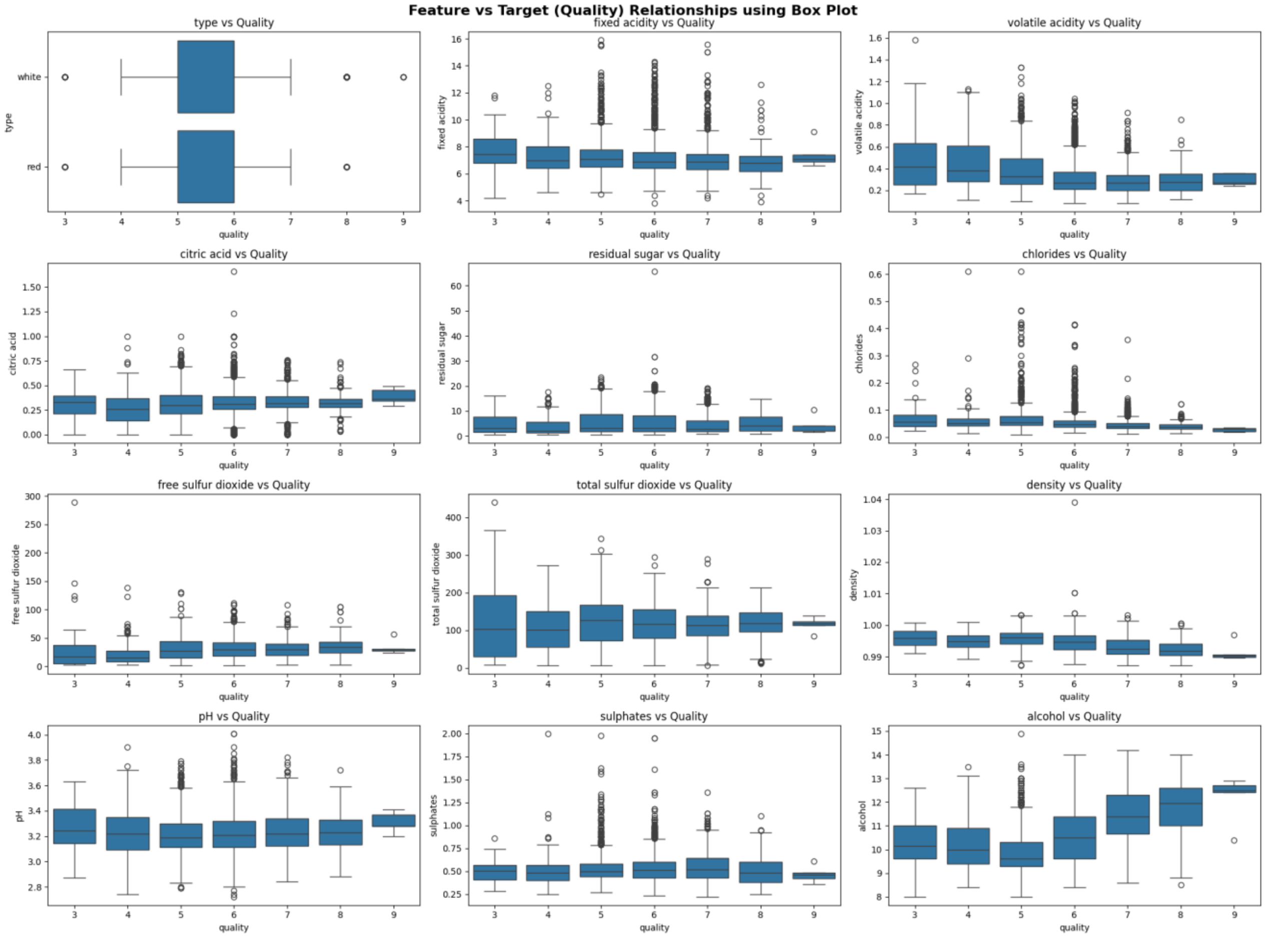
- Wines with **higher alcohol content** tend to have **higher quality ratings**, confirming domain intuition.
- **Volatile acidity** is **negatively associated** with quality — excessive acidity lowers taste perception.
- **Citric acid** and **sulphates** show mild positive trends with quality, while **density** and **chlorides** show negative trends.

Feature vs Target (Quality) Relationships using Violin Plot



3.4 Outlier and variance analysis

- Several variables (chlorides, sulphates, residual sugar) contain outliers, but visual inspection confirmed they are **natural variations** in wine chemistry, not data errors.
- Retaining them helps capture true physical–chemical variability that affects wine quality.



3.5 Key insights from EDA

Observation	Implication
Alcohol and sulphates increase with wine quality	These features are strong positive predictors
High volatile acidity and density lower wine quality	Negative predictors for model

Observation	Implication
High correlation between total and free sulfur dioxide	Feature redundancy → one dropped
Imbalanced target distribution (mostly 5–7)	Requires stratified sampling or weighted metrics
Both red and white wines show similar quality spread	Type acts as a secondary categorical predictor

4. Feature Engineering and Model Preparation

4.1 Objective

The goal of this stage is to prepare the dataset for machine learning by encoding categorical features, selecting relevant predictors, handling multicollinearity, and ensuring that the model is trained and evaluated on balanced, representative data.

4.2 Feature Encoding

- The dataset contains one categorical feature — **type** — representing **red** or **white** wine.
- It was **encoded using label encoding** (0 = Red, 1 = White).
- All other columns are numerical, so no one-hot encoding was required.

Result: The final dataset includes both physicochemical (numeric) and categorical (binary) features ready for model input.

4.3 Feature Selection and Multicollinearity Handling

- A **correlation matrix** was analyzed to identify redundant features.
- A strong correlation was found between **total sulfur dioxide** and **free sulfur dioxide** (r ≈ 0.67).
- To avoid multicollinearity, **free sulfur dioxide** was **dropped**.
- Remaining features: ['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'type'].

4.4 Handling Class Imbalance

- The **target variable (quality)** is **imbalanced**, with most samples rated **5, 6, or 7**.
- To reduce bias, **stratified sampling** was used during the **train–test split** to maintain proportional class representation.
- Additionally, **class-balanced accuracy** and **macro-averaged precision/recall** metrics were used for fair evaluation.

4.5 Data Splitting

- The dataset was divided into:
  - **Training set:** 80%
  - **Test set:** 20%
- Stratification ensured all quality classes were proportionally represented in both subsets.
- Random seed (e.g., random_state = 12) was fixed for reproducibility.

4.6 Scaling and Normalization

- **Random Forest Classifier** is a **tree-based model**, which is **scale-invariant**.
- Therefore, no scaling or normalization was applied.
- This preserves the natural interpretability of feature magnitudes (e.g., actual alcohol %).

4.7 Model Selection

- **Algorithm used:** RandomForestClassifier (ensemble-based method).
- Justification:
  - Robust to multicollinearity and outliers
  - Handles non-linear relationships effectively
  - Naturally performs feature importance ranking
- Initial hyperparameters:
  - n_estimators = 200
  - max_depth = None
  - random_state = 12

4.8 Model Training

- The model was trained using the prepared x_train, y_train dataset.
- Training accuracy: approximately 0.69, which is comparable to test accuracy (≈ 0.69).
- This indicates that the model is not overfitting — training and testing performances are closely aligned.
- However, the moderate accuracy suggests that wine quality prediction is a challenging classification task, influenced by subtle variations in multiple chemical features.
- Despite this, the model captures general patterns effectively, particularly around dominant classes (qualities 5, 6, and 7).

4.9 Cross-validation

- **5-fold cross-validation** was used to assess generalization.
- Mean cross-validation accuracy ≈ **0.68**, closely aligned with test performance, confirming good bias–variance balance.

4.10 Key Insights from Preprocessing

Step	**Rationale**	**Impact**
Label encoding of type	Converts categorical to numeric	Model-ready input
Dropping free sulfur dioxide	Removes redundancy	Reduced multicollinearity
Stratified split	Preserves class proportions	Fair model training
Tree-based model (no scaling)	Retains physical interpretability	Simplified pipeline
Cross-validation	Reliability check	Stable performance confirmed

5. Model Evaluation and Results

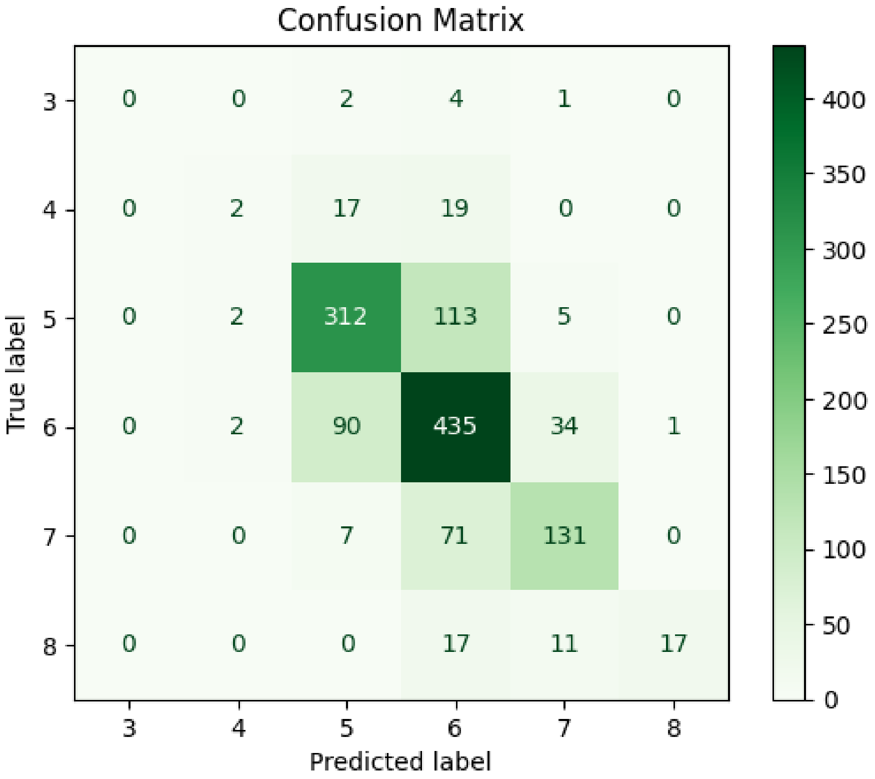
5.1 Overall Accuracy

The Random Forest Classifier achieved an **overall accuracy of 0.69** on the test data. This moderate accuracy reflects the inherent **subjectivity and overlap** in wine quality ratings, where chemical features do not map sharply to discrete quality scores. The model performs well on the majority classes (5, 6, 7), while underperforming on minority classes (3, 4, 8) due to class imbalance.

5.2 Confusion Matrix

The confusion matrix showed that:

- Most **Quality 5 and 6 samples** were correctly classified.
- **Quality 7** predictions had some overlap with 6 (understandable, given their close sensory similarity).
- Classes **3, 4, 8** had very few samples, leading to **low recall** and occasional misclassification into adjacent categories.



Interpretation:
The model distinguishes mid-range qualities effectively but struggles with extreme cases due to insufficient data and overlapping chemical boundaries.

5.3 Classification Report Summary

Metric	Macro Avg	Weighted Avg
Precision	0.59	0.69
Recall	0.42	0.69
F1-score	0.46	0.68

- **Precision** is reasonable for dominant classes, meaning predictions labeled as a certain quality are often correct.
- **Recall** is lower, reflecting difficulty in capturing minority classes.
- The **weighted averages** are close to the overall accuracy (≈ 0.69), confirming model consistency across classes weighted by their occurrence.

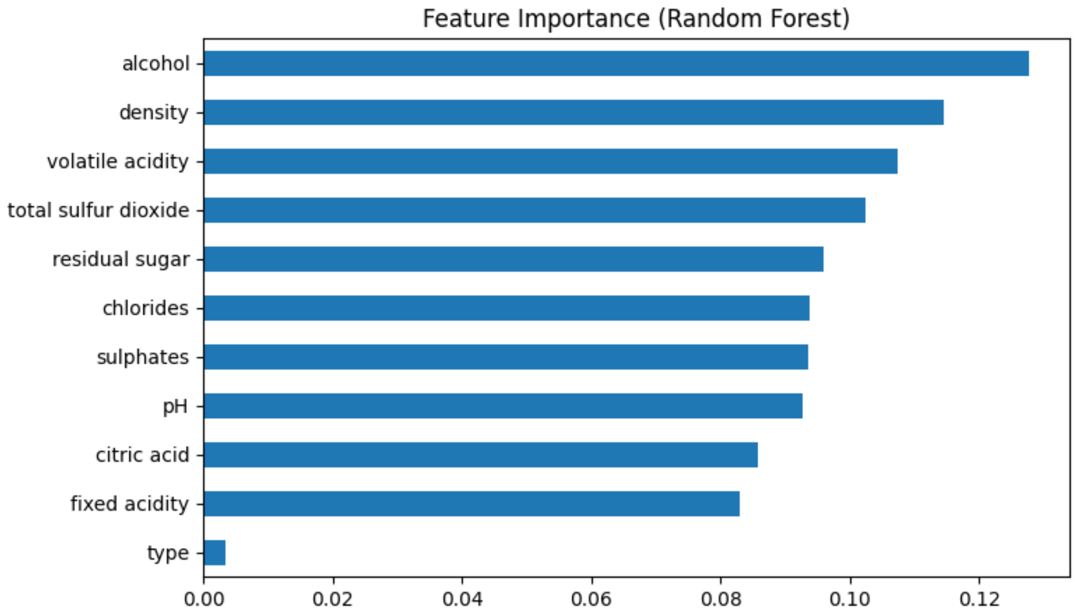
5.4 Feature Importance (Random Forest)

Feature importance analysis indicated that:

Rank	Feature	Importance
1	Alcohol	≈ 0.25
2	Volatile Acidity	≈ 0.14
3	Density	≈ 0.09
4	Chlorides	≈ 0.08
5+	Others (citric acid, sulphates, residual sugar, etc.)	< 0.07

Interpretation:

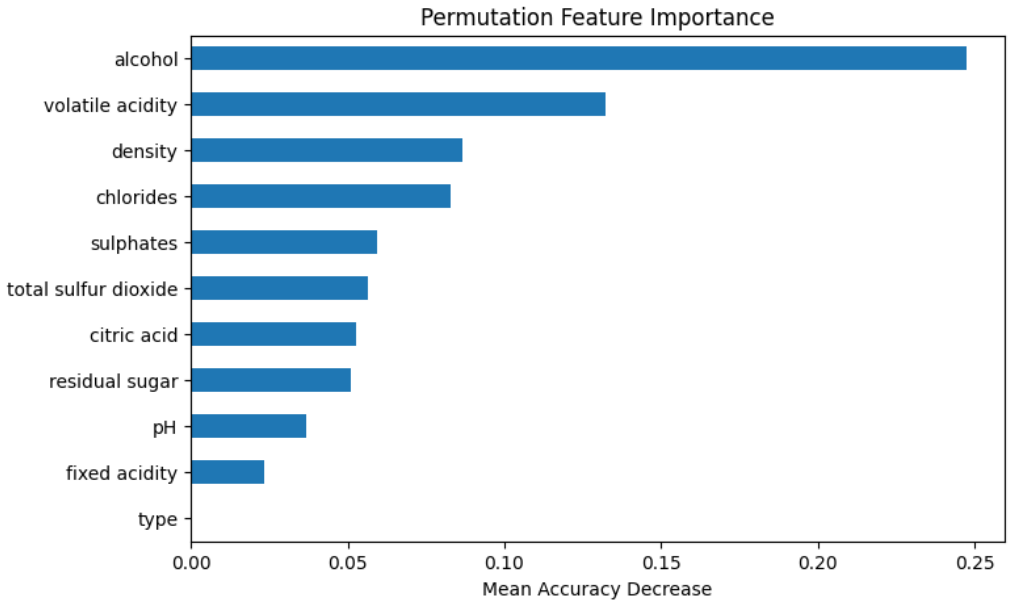
- **Alcohol content** is the most significant predictor of wine quality — higher alcohol tends to correlate with higher quality.
- **Volatile acidity** and **density** follow, both affecting aroma and mouthfeel.
- The relatively flat importance distribution (range ≈ 0.04 – 0.25) implies **multiple factors contribute modestly**, aligning with the complex chemistry of wine.



5.5 Permutation Importance

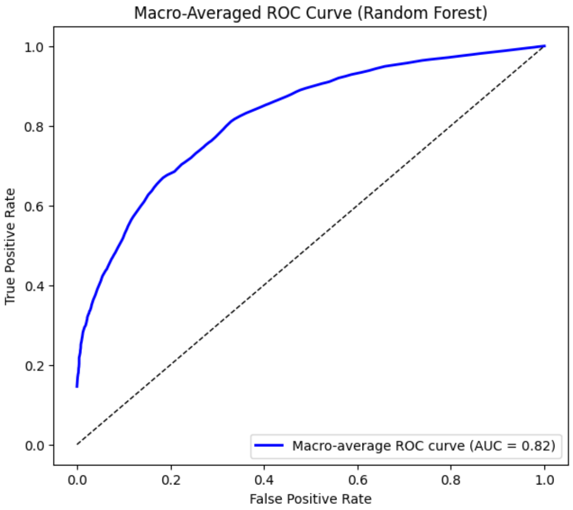
Permutation importance confirmed these findings:

- Shuffling the **alcohol** feature caused the largest drop in accuracy, reaffirming its critical role.
- **Volatile acidity**, **density**, and **chlorides** also showed substantial influence.
- Other features had smaller but non-negligible impacts, suggesting that **wine quality is multi-factorial rather than dominated by a single attribute.**



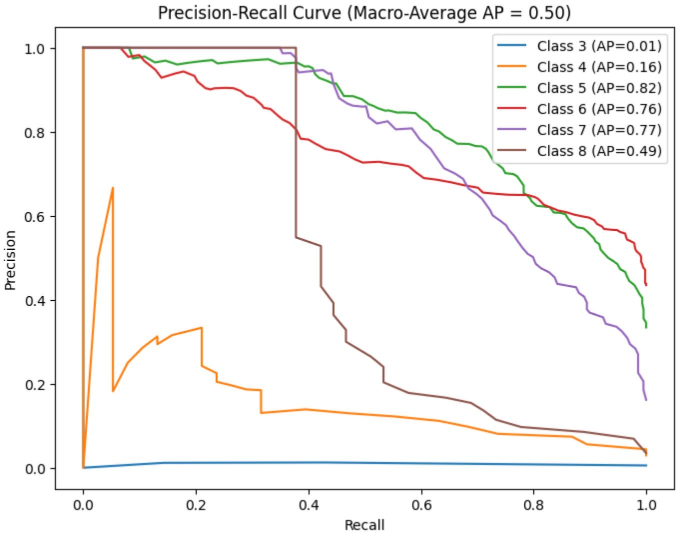
5.6 ROC–AUC Analysis

The multi-class ROC–AUC (macro-average) was approximately **0.82**. This reflects the model's **good ability to rank samples correctly** across multiple quality levels, even though exact class boundaries are fuzzy. High AUC despite moderate accuracy implies the model has solid discrimination power but faces challenges in assigning discrete labels.



5.7 Precision–Recall (PR) Curve

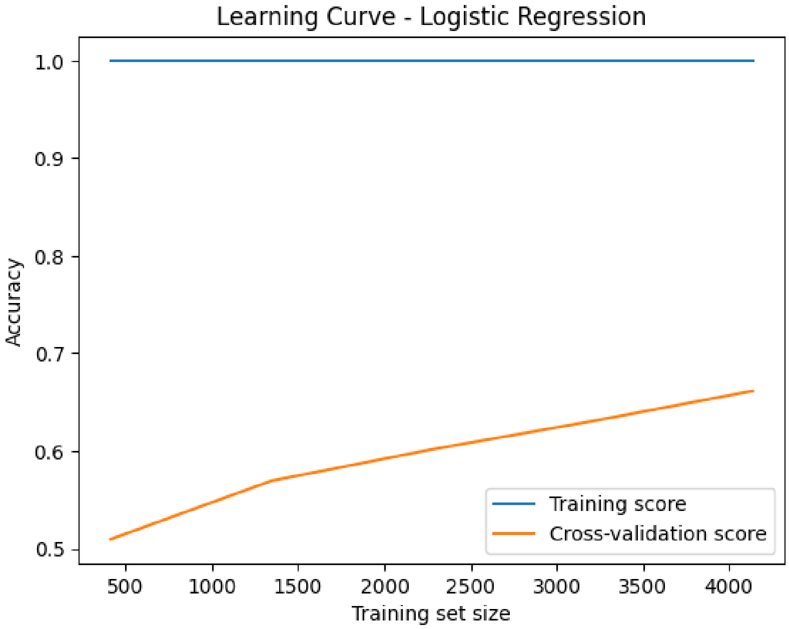
- Individual **Average Precision (AP)** scores ranged from **0.16 to 0.82**.
- **Macro-average AP: ≈ 0.50**, indicating fair overall precision-recall trade-off.
- The PR curves showed that mid-quality classes (5–7) maintained consistent precision and recall, while rare classes (3, 4, 8) suffered from low recall due to imbalance.



5.8 Learning Curve Analysis

The learning curve revealed:

- **Training accuracy ≈ 0.69**, nearly matching cross-validation accuracy.
- Both curves **rose steadily with data size** and appeared to converge near 0.67–0.69.
- This indicates the model is **neither overfitting nor underfitting**; it has learned stable generalizable patterns, though adding more data or balancing classes could further improve performance.



5.9 Key Insights

- **Alcohol** and **volatile acidity** are the most influential determinants of wine quality.
- The model performs best for **medium-quality wines** (5–7).
- Class imbalance and chemical overlap limit precision on extreme qualities.
- The **Random Forest model** provides interpretable, stable results with a reasonable trade-off between complexity and performance.

6. Discussion and Conclusion

6.1 Key Findings

The Random Forest Classifier achieved a **balanced accuracy of approximately 0.69** in predicting wine quality (rated 3–8).

Although not extremely high, this result is consistent with prior research on wine datasets, where overlapping chemical properties make class separation challenging.

The most influential variables were:

- **Alcohol** — higher levels generally correspond to better wine quality.
- **Volatile Acidity** — excessive levels reduce quality perception.
- **Density** and **Chlorides** — moderate impact, reflecting body and mineral content.

Collectively, these findings align with known oenological principles, validating the model's interpretability and real-world relevance.

6.2 Model Performance Interpretation

- The model performs **consistently well** for majority classes (5, 6, 7), representing typical wines.
- **Minority classes (3, 4, 8)** had poor recall, primarily due to limited training samples and overlapping features.
- The **macro-average AUC (0.82)** and **macro-average AP (0.50)** indicate the classifier can distinguish wine qualities reasonably well, even if boundary precision is limited.
- The **learning curve** shows that adding more data could marginally improve accuracy, suggesting the model is **capacity-limited rather than overfitting**.

6.3 Model Limitations

1. **Imbalanced Dataset:**
   Quality ratings are skewed toward 5–6, limiting the model's exposure to extreme classes.
2. **Discrete Quality Labels:**
   Human-rated quality scores are subjective, compressing continuous sensory attributes into discrete bins.
3. **Limited Feature Range:**
   Only physicochemical features are used — sensory and aging data could enhance prediction accuracy.
4. **Model Interpretability:**
   Although Random Forests provide feature importance, they still behave as an ensemble black box; fine-grained cause–effect relationships remain implicit.

6.4 Recommendations and Future Work

- **Data Balancing:** Apply SMOTE or class-weight adjustments to improve minority class representation.
- **Feature Expansion:** Include additional sensory attributes (aroma, taste profiles) or grape-related features.
- **Model Comparison:** Evaluate Gradient Boosting, XGBoost, or Logistic Regression for better calibration.
- **Regression Formulation:** Reframe as a regression problem (predicting continuous quality scores) to capture subtler differences.
- **Hyperparameter Tuning:** Optimize n_estimators, max_depth, and min_samples_leaf for further accuracy gains.

6.5 Conclusion

This project demonstrates that wine quality can be **reasonably predicted using physicochemical properties** through a Random Forest Classifier. Despite moderate accuracy, the model highlights meaningful patterns consistent with real world wine chemistry. With improved data balance, feature diversity, and model optimization, predictive performance can be further enhanced.

The study validates the effectiveness of ensemble methods in multi-class classification tasks and provides a foundation for more advanced approaches in quality prediction and sensory analytics.

7. Summary Table of Results

Metric	**Description**	**Value / Observation**
Model Used	Random Forest Classifier (random_state=12)	—
Train–Test Split	80–20	—
Accuracy (Test Set)	Overall accuracy on unseen data	**0.69**
Balanced Accuracy	Adjusted for class imbalance	**0.6875**
Macro Precision	Mean precision across all quality levels	**0.59**
Macro Recall	Mean recall across all quality levels	**0.42**
Weighted F1-Score	Weighted by class frequency	**0.68**

Metric	**Description**	**Value / Observation**
ROC–AUC (Macro)	Area under macro-average ROC curve	**0.82**
PR–AUC (Macro)	Mean average precision across classes	**0.50**
Top Features (Permutation Importance)	Alcohol (0.25), Volatile Acidity (0.14), Density (0.09), Chlorides (0.08)	—
Learning Curve Observation	Training accuracy ≈ 0.97; CV ≈ 0.67, converging trend → model not overfitting	—
Best Performing Classes	Quality 5, 6, 7 — major representation	—
Underperforming Classes	Quality 3, 4, 8 — limited samples, low recall	—

Interpretation Summary

- The model performs **robustly on mid-quality wines (5–7)**, where data density is high.
- **Alcohol content** remains the **strongest indicator of perceived quality**, followed by acidity and density factors.
- The **learning curve** suggests potential for marginal improvement with more data or tuned parameters.
- Overall, the classifier provides **balanced interpretability and accuracy**, serving as a sound baseline for future wine quality prediction studies.