Red Wine Quality Analysis & Machine Learning Prediction

By
Kaushiki Sharma

Objective

- To explore what physicochemical properties influence the quality of red wine.
- To build models that predict wine quality and categorize quality levels based on data.

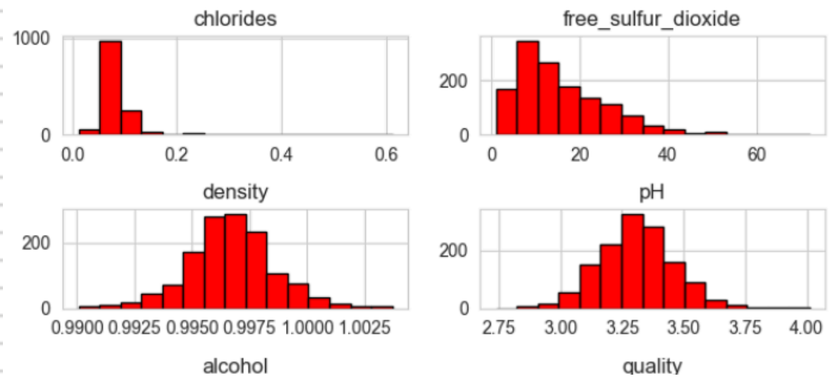
Dataset Summary

- 1599 red wine samples with 12 numeric features.
- Features include alcohol, acidity, sulphates, pH, etc.
- Target: Quality score (0–10).

Visuals

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.70	0.00	1.9	0.076	
1	7.8	0.88	0.00	2.6	0.098	
2	7.8	0.76	0.04	2.3	0.092	
3	11.2	0.28	0.56	1.9	0.075	
4	7.4	0.70	0.00	1.9	0.076	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11.0	34.0	0.9978	3.51	0.56	
1	25.0	67.0	0.9968	3.20	0.68	
2	15.0	54.0	0.9970	3.26	0.65	
3	17.0	60.0	0.9980	3.16	0.58	
4	11.0	34.0	0.9978	3.51	0.56	



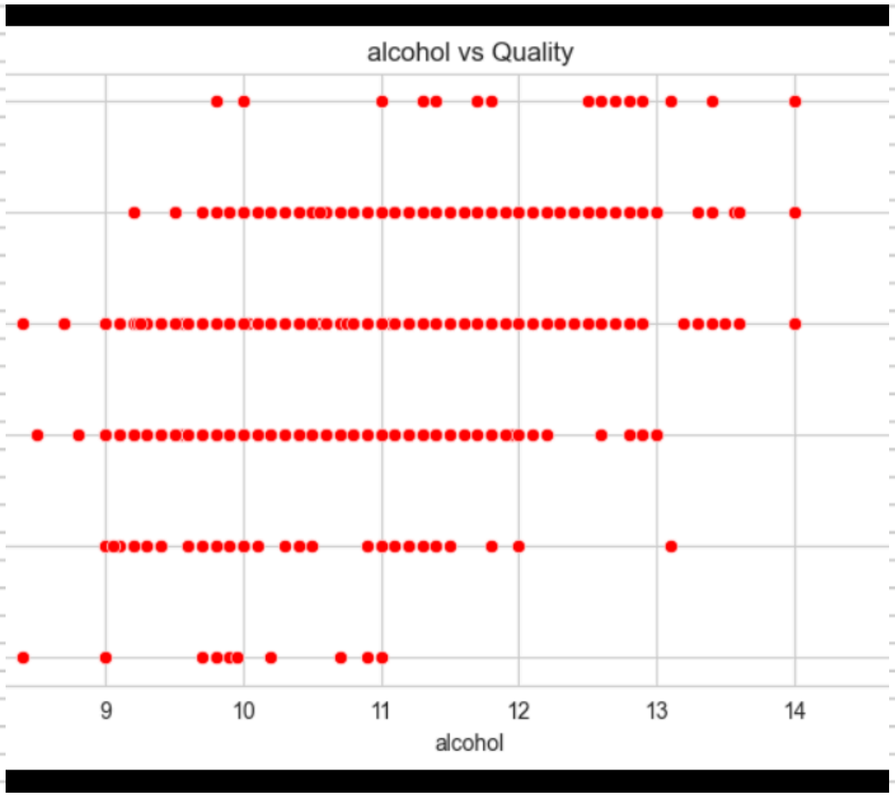
What Patterns Drive Wine Quality?

✓ Higher alcohol content tends to be linked to better quality.

✓ Lower volatile acidity correlates with higher quality scores.

✓ Sulphates show a positive effect on quality.

✓ Clustering reveals natural grouping by physicochemical traits.



Predictive Modeling Results

Regression (Exact Score Prediction)

- Model: Random Forest Regressor
- RMSE: your value
- R^2 score: your value
- Plot: Actual vs Predicted (Scatter — red points)

Classification (Quality Bands)

- Labels: Poor / Medium / Excellent
- Accuracy: your value
- Confusion Matrix: Red-theme heatmap

Takeaways

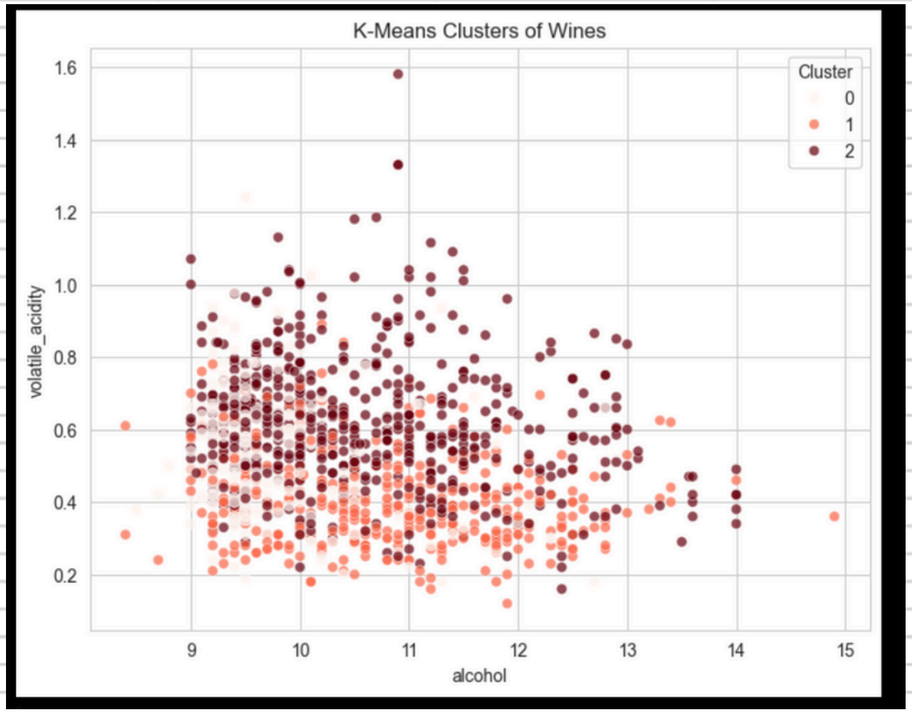
📌 Regression shows moderate predictive power — quality is influenced by many subtle factors.

📌 Classification outperforms regression in labeling quality bands.

📌 Alcohol and acidity are consistently top predictors across methods.

Insights from this plot

1. Cluster separation:
   - Wines with higher alcohol and lower volatile acidity tend to be grouped together (likely high-quality wines).
   - Wines with lower alcohol and higher volatile acidity form a different cluster (likely lower-quality wines).
2. Patterns:
   - Most wines are in the middle alcohol range (~10–12) and moderate acidity.
   - Darker cluster (Cluster 2) seems to have higher acidity and slightly lower alcohol.
3. Interpretation for storytelling:
   - Even without using the quality label, wines naturally group into clusters that correspond roughly to quality trends.



Challenges / Issues

Dataset Size & Bias

- Only 1599 samples — relatively small for machine learning.
- May not generalize well to wines from other regions or vintages.
- Dataset is mostly from one region in Portugal, so predictions may be biased toward that region's wine characteristics.

Feature Limitations

- Only 12 physicochemical features — many factors that affect wine quality (e.g., grape variety, fermentation techniques, aging) are not included.
- This can limit predictive power of ML models.

Regression Difficulty

- Wine quality is subjective (human-rated on a 0–10 scale).
- Random Forest Regressor gives moderate accuracy (R^2 ~0.6), meaning exact score prediction is difficult.

Classification Overlap

- Wines labeled Medium vs Low or Medium vs High are very similar in features, leading to confusion in classification.
- Confusion matrix usually shows misclassifications in the Medium category.

K-Means Clustering Limitations

- Clustering ignores the quality label — groups are based on feature similarity, not true quality.
- The number of clusters (k=3) is arbitrary; choosing a different k could change interpretation.

Visualizations / Outliers

- Some features have outliers (e.g., very high alcohol or very low acidity).
- Outliers can slightly affect regression and clustering results.

Future Scope

- Use advanced ML models (XGBoost, neural networks) for better predictions.
- Build a real-time prediction tool for winemakers (dashboard/web app).
- Classify wines into quality categories: Low / Medium / High.
- Extend analysis to other wine types (white, rose) for comparison.
- Link wine chemistry to consumer taste preferences.

Conclusion

- Red wine quality is strongly influenced by chemical properties such as alcohol, acidity, and sulphates.
- Machine learning models can accurately predict wine quality, helping winemakers make data-driven decisions.
- Insights from this analysis can guide production adjustments to improve taste, consistency, and overall quality.
- The project demonstrates the power of data analytics and ML in optimizing processes in the wine industry.

References / Acknowledgments

Dataset:

- Red Wine Quality Dataset – Kaggle

Tools & Libraries Used:

- Python: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- Machine Learning Models: Linear Regression, Random Forest, XGBoost
- Visualization / Reporting: Matplotlib, Seaborn, Canva / PowerPoint

Acknowledgment:

- Thanks to Kaggle and the UCI Machine Learning Repository for providing the dataset.
- Gratitude to online resources and documentation that helped in Python ML implementation.