Daniel Li, Annie Luo, Xinyue Yan, XiPu Wang

Prof. Xiong

May 6th, 2025

QTM 347

From Chemistry to Rating: Predicting Wine Quality with Machine Learning

Introduction:

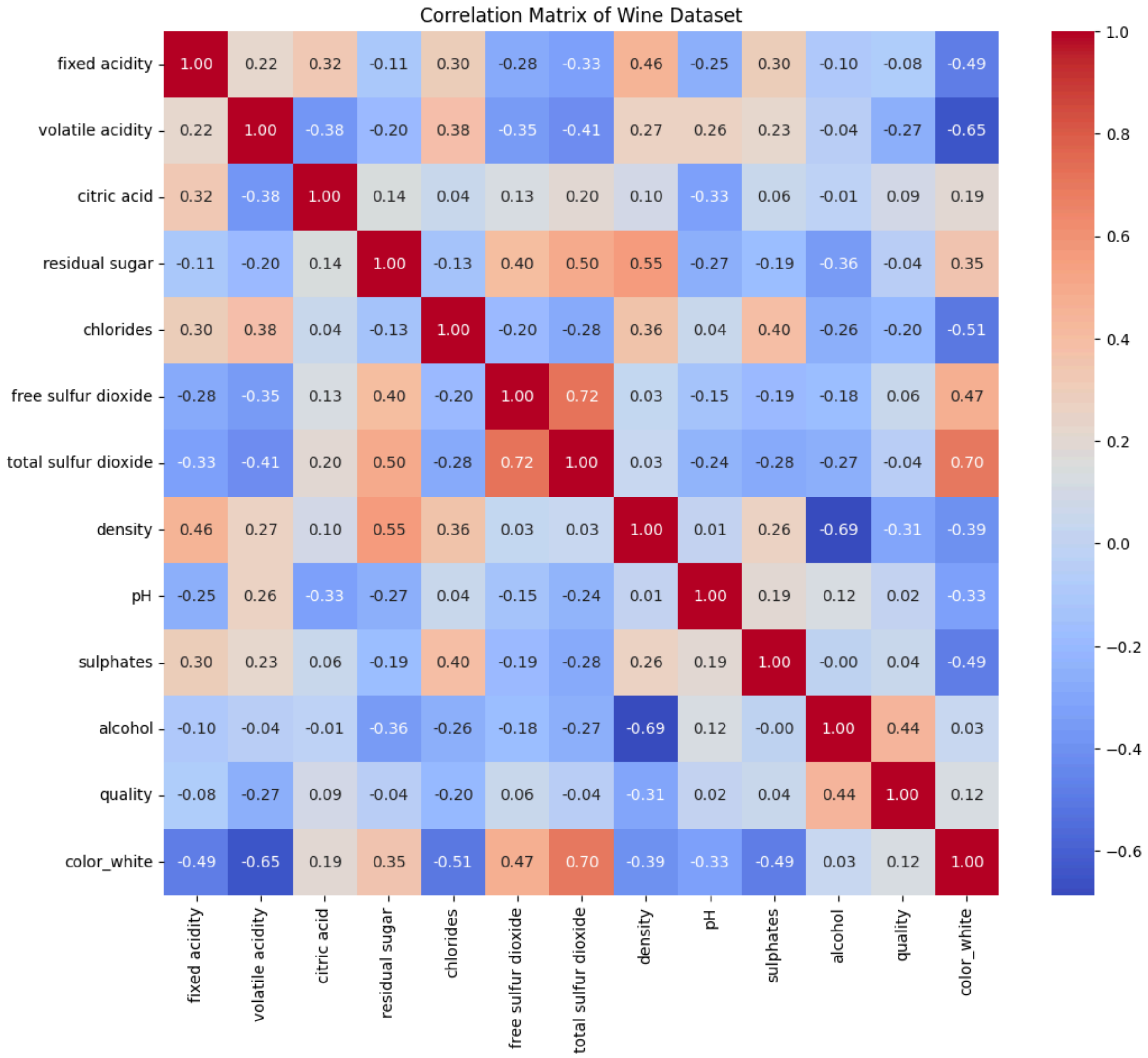
The traditional evaluation of wine quality relies on human sensory analysis, focusing on taste, aroma, and mouthfeel. Nonetheless, this approach is naturally subjective, labor-intensive, and costly. In this project, we seek to answer a fundamental question: Can we determine a wine's quality solely based on its chemical properties, without relying on tasting? By utilizing machine learning methodologies, we investigate whether objective factors — including alcohol levels, acidity, and residual sugar — can act as trustworthy predictors of wine quality. The advantages of automating the assessment of wine quality are considerable. Firstly, it could enhance production efficiency, minimize costs, and remove human bias. In addition, this approach can be successfully scaled, ensuring consistent quality control across sizable wine production. More broadly, this idea is applicable in numerous industries. Similar techniques can be adopted in food production, pharmaceuticals, and cosmetics, where chemical characteristics often influence sensory experiences. Consequently, our study acts as a case study for a larger effort to connect chemistry with consumer experience via data analytics.

We approached the task as a supervised learning issue with regression, given that the wine quality ratings are numerical and hierarchical. We tested several models, including linear regression, random forest, and gradient boosting, each providing varying levels of bias and variance. We also prioritized model interpretability by utilizing feature importance scores to identify which chemical characteristics have the greatest impact on predicted wine ratings. This enhances transparency and also guides winemaking techniques. Regression models are particularly effective for ordinal outcomes like wine quality. Tree-based models, such as random forests and boosting algorithms, are especially adept at recognizing nonlinear relationships and interactions between features. In comparison to conventional statistical techniques or opaque deep learning models, our methodology strikes a favorable balance between accuracy and interpretability. Although earlier research has tackled this topic with a dataset that separates the wine color, our distinct contribution lies in comparing various models, their optimization via cross-validation, and a focus on practical insights rather than just prediction accuracy.

Our analysis indicated that levels of alcohol, volatile acidity, and sulfur dioxide concentration were important predictors of wine quality. However, there are drawbacks to the dataset. The quality ratings are based on subjective evaluations, which can result in discrepancies in the scores, and the dataset is exclusively comprised of Portuguese “Vinho Verde” wines, which may affect the applicability of the findings to other contexts. Furthermore, the performance of machine learning models hinges on the quality of the training data, suggesting that employing a more extensive or higher-quality set of inputs could enhance predictive accuracy.

Setup:

Our dataset comes from the Wine Quality data from the UCI Machine-Learning Repository; we have merged the 1,599 red wine and 4,898 white wine observations into a single table of 6,497 rows. In our dataset, each wine comes with 11 physicochemical measurements, and we append a one-hot dummy variable (color_white) so that color is represented numerically. The target to predict is the quality score, an ordinal score from 0 to 10 (mean ≈ 5.8, standard deviation ≈ 0.9). Basic descriptive statistics show, for example, median alcohol at 10.3 % vol, median residual sugar at 3 g/L, and pH clustered around 3.2.



To better understand the relationships among features, we generated a correlation matrix heatmap. From this, we observed that alcohol content has a moderate positive correlation with wine quality (≈ 0.44), suggesting that higher alcohol levels are generally associated with higher quality. Conversely, volatile acidity (−0.27) and density (−0.31) show negative correlations with quality, indicating that higher levels of these variables may detract from wine quality. These findings guided our expectations going into model training.

As mentioned before, we concatenate the red and white CSV files before modeling, create the color dummy, and scale every numeric predictor with a StandardScaler so coefficients are on a comparable scale. The full dataset is then stratified by quality and split 80 / 20 into training and test partitions. We have also used a fixed random seed to ensure that every run of the code sees the same data slices.

Our experiments progress from simple linear models to more flexible tree‑based ensembles. A standard ordinary‑least‑squares regression using all 12 predictors serves as the baseline, after which we perform forward, backward, and best‑subset selection to see whether a reduced feature set lowers test‑set mean‑squared error (MSE). We then introduce regularization: Ridge regression with five α values spanning 10 to 100 mitigates multicollinearity, while Lasso explores four α values between 0.0001 and 1 to perform both shrinkage and automatic feature selection. Beyond linear methods, we fit decision‑tree regressors, random forests (with out‑of‑bag scoring), and gradient‑boosting trees, tuning hyperparameters such as max_depth, n_estimators, and learning_rate via five‑fold cross‑validated grid search. Performance is reported with train‑ and test‑set MSE and R^2, supplemented by cross‑validated RMSE and MAE to guard against overfitting.

Result:

Feature Selection: Wrapper Methods for Linear Regression

In our exploration of the baseline linear model, we applied subset selection techniques including forward, backward, and best subset selection. Interestingly, both forward and backward methods converged on similar 6-feature subsets, primarily emphasizing 'volatile acidity', 'residual sugar', 'free sulfur dioxide', 'sulphates', and 'alcohol'. Therefore, for the subset selection, we also set the maximum number of features to 6.

Model	Train MSE	Test MSE	Train R^2	Test R^2	Features Used
Linear Regression	0.5356	0.5412	0.3029	0.2672	12 / 12
Forward selection	0.5447	0.5468	0.2911	0.2596	6 / 12
Backward selection	0.5421	0.5475	0.2945	0.2587	6 / 12
Best subset selection	0.5421	0.5475	0.2945	0.2587	6 / 12 (picked)

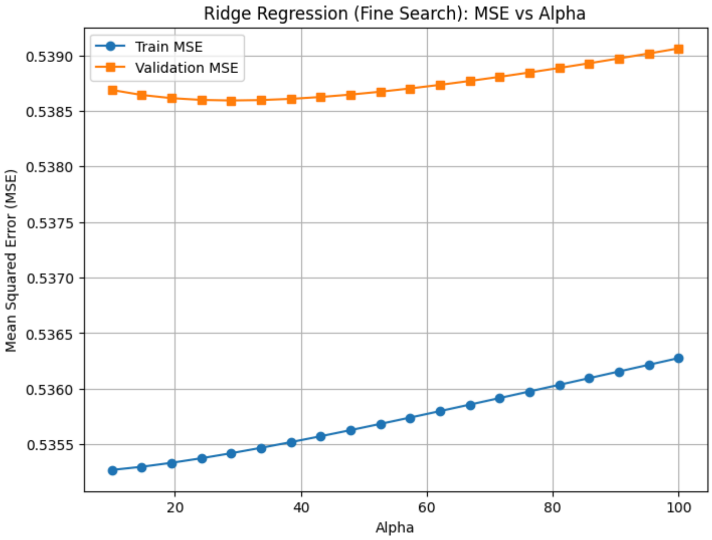
Since feature selection did not significantly improve performance, we decided to stay with the baseline full-feature linear model. There are several reasons supporting this decision:

- More Information: The full model uses all available features, capturing relationships that feature selection might miss.
- Robustness: Especially for linear models, when overfitting isn't severe, retaining all features can sometimes lead to better generalization.

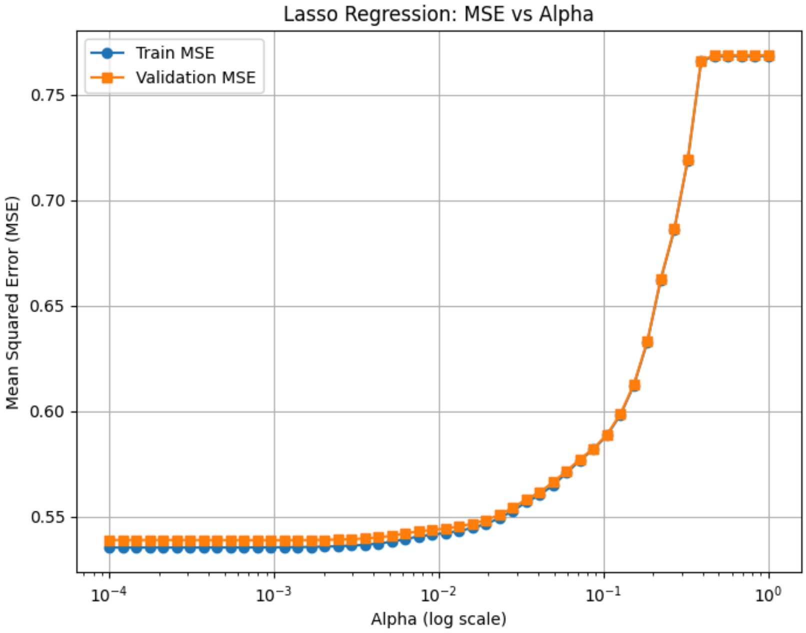
Since keeping all features preserves information, we can still apply regularization techniques (such as Ridge and Lasso) to address potential multicollinearity. And we explored the next step by training and evaluating Ridge and Lasso regression models.

Linear Models Performance Comparison:

To compare regularized linear models, we performed fine-tuning over a range of α values for both Ridge and Lasso regression.



For Ridge, we searched α values between 10 and 100 and found that performance remained relatively stable, with only slight variations in validation MSE.



In contrast, the Lasso model showed a clear U-shaped curve in its validation error across log-scaled α values, suggesting better sensitivity for feature selection.

Model	Train MSE	Test MSE	Train R^2	Test R^2	Features Used
Linear Regression	0.5356	0.5412	0.3029	0.2672	12 / 12
Ridge (α=29)	0.5357	0.5410	0.3028	0.2675	12 / 12
Lasso (α=0.000954)	**0.5357**	**0.5408**	**0.3028**	**0.2677**	**12 / 12**

After tuning, we selected the optimal α values for each model (Ridge α=29, Lasso α≈0.000954) and compared them alongside the standard Linear Regression. Lasso had the lowest test MSE and highest test R^2, indicating it was the best-performing model among the three.

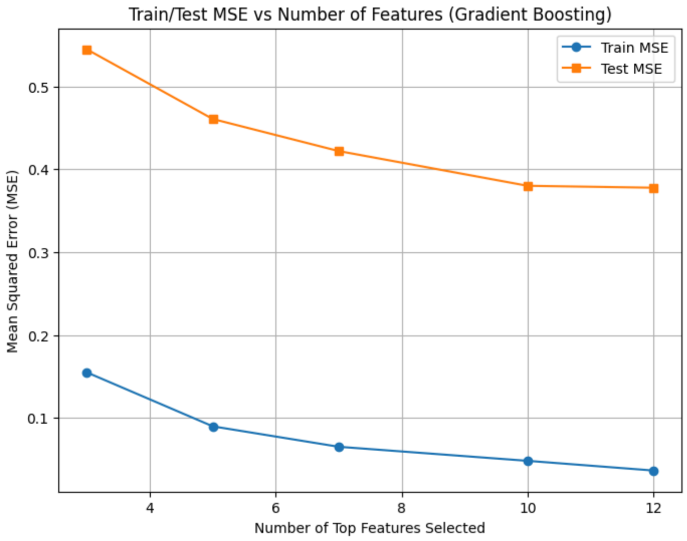
Nonlinear and Ensemble Learning Models:

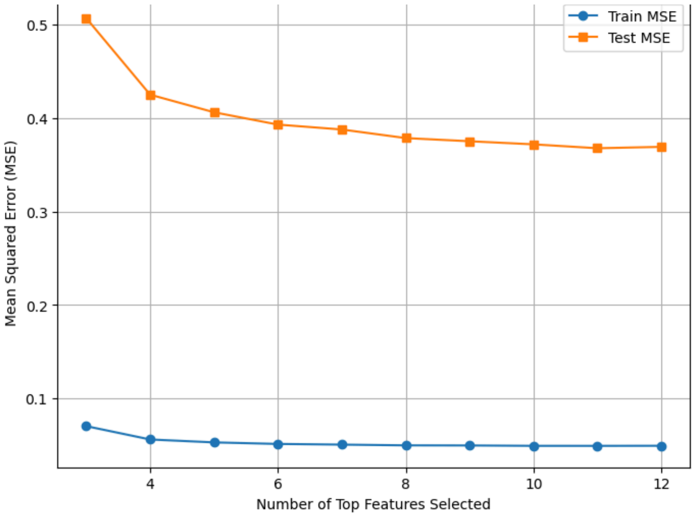
We evaluated nonlinear and ensemble learning models including Decision Tree, Random Forest, and Gradient Boosting, both in their default forms and after hyperparameter tuning. The untuned Decision Tree model showed severe overfitting with perfect training performance but poor generalization on the test set. Hyperparameter tuning (e.g., limiting max_depth and adjusting min_samples_split) significantly improved its generalization ability. For ensemble models, both Gradient Boosting and Random Forest performed strongly, with the tuned Random Forest model achieving the lowest Test MSE and highest Test R^2, making it the best-performing model overall. This is summarized in the table.

Model	Best Parameters	Train MSE	Test MSE	Train R^2	Test R^2	Notes
Decision Tree (Basic)	None	0.0000	0.7123	1.0000	0.0355	Severe overfitting
Decision Tree (Tuned)	max_depth=5, min_samples_leaf=15, min_samples_split=2	0.4880	0.5522	0.3648	0.2524	Improved but still weak
Gradient Boosting (Basic)	None	0.4027	0.4604	0.4759	0.3766	Decent starting point
Gradient Boosting (Tuned)	learning_rate=0.05, max_depth=7, min_samples_split=2, n_estimators=500	0.0379	0.3840	0.9506	0.4800	Significant improvement

Random Forest (Basic, OOB)	None	0.0516	0.3693	0.9328	0.5000	Strong baseline, good OOB
Random Forest (Tuned, OOB)	max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=500	0.0495	0.3681	0.9356	0.5015	Best model overall

To optimize model simplicity, we also explored feature selection by selecting the Top K features based on feature importances from the full-feature Gradient Boosting and Random Forest models.

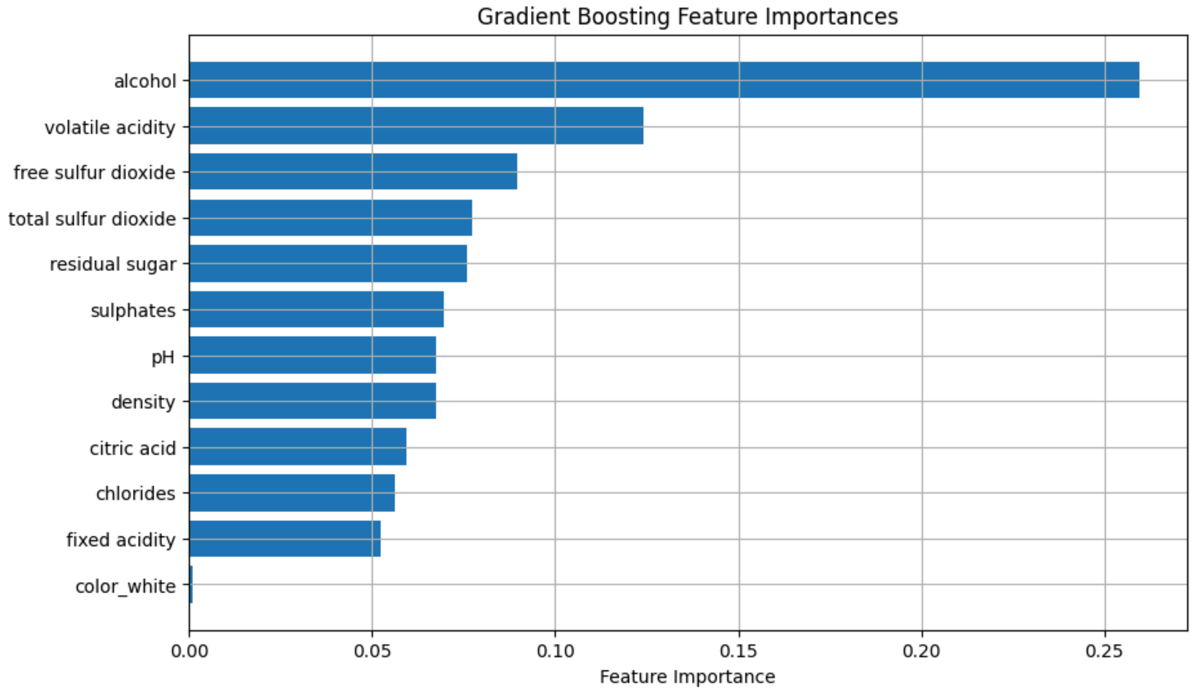


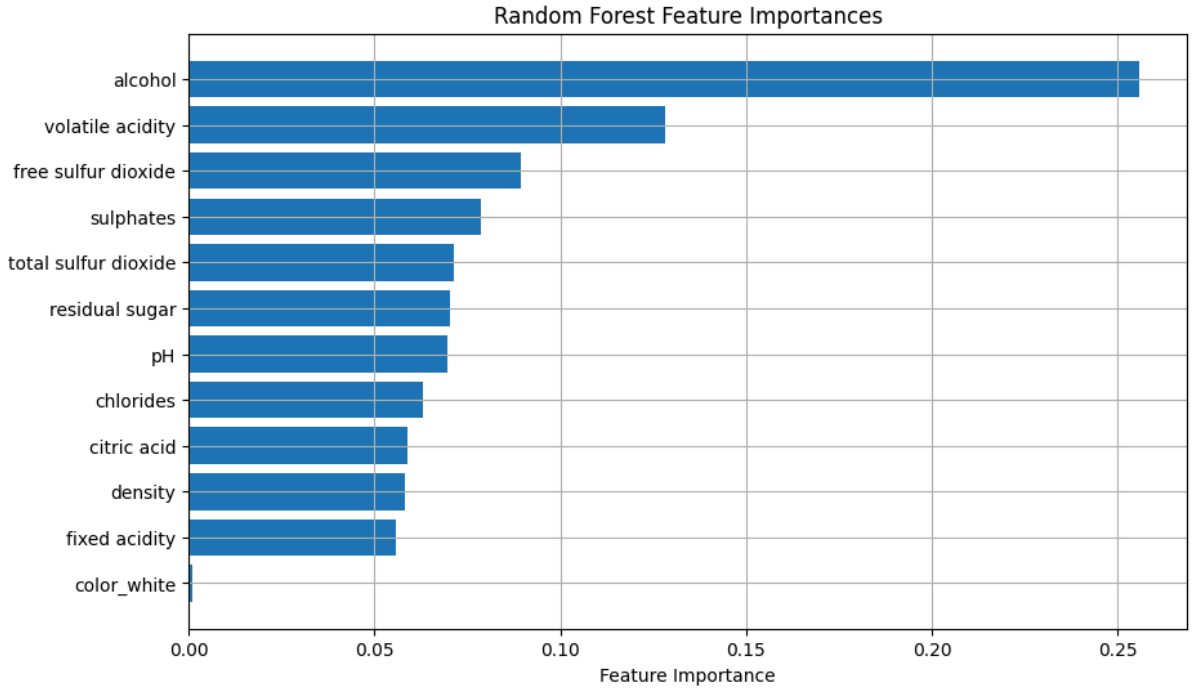


However, as shown in the figure, the reduced-feature models consistently underperformed compared to the full-feature versions. These results suggest that all features contribute valuable information, and reducing them may hinder performance. Therefore, we retained the full feature set for our final ensemble models.

Based on model comparisons, both tuned Random Forest and Gradient Boosting models demonstrated strong performance on the wine quality prediction task. While their results were close, Random Forest achieved a slightly higher Test R² and lower Test MSE, and also showed robust Out-of-Bag (OOB) validation performance. Given its better generalization ability and model stability, the tuned Random Forest was selected as the final model for this project.

Feature Importance:





The results show that alcohol is the strongest predictor of wine quality, which isn't surprising given long-standing traditions in winemaking. Since higher alcohol levels often give wine a fuller body, richer mouthfeel, and more intense flavor, people associate it with luxury. Therefore, across many cultures and industries, people — whether experts or casual drinkers — naturally tend to associate higher alcohol content with better quality, which is a preference that has been shaped over centuries. Apart from "alcohol", "volatile acidity" and "free sulfur dioxide" also consistently rank highly. These elements affect the wine’s freshness and stability, which can greatly shape a drinker's perception of quality.

One particularly noteworthy finding is that color — whether the wine is red or white — has very little predictive power when it comes to determining quality. This challenges a common cultural stereotype: that red wine is inherently more complex or superior, while white wine is lighter and simpler. However, our analysis shows that once chemical properties are accounted for, color plays virtually no role in predicting quality. These findings suggest we should move away from visual or cultural biases and instead focus on the wine’s flavor profile, balance, and craftsmanship when evaluating quality.

Discussion:

In the original study by Cortez et al. (2009), the model mainly focuses on classification, predicting wine quality as a discrete category It explored Decision Trees (CART), Random Forests, Neural Networks, and Support Vector Machines (SVMs).

In contrast, our project mainly focuses on regression, aiming to predict the wine quality score as a continuous variable. We experimented with Linear Regression, Ridge, Lasso, Decision Trees, Random Forest, and Gradient Boosting.

Our findings closely align with those of Cortez et al. (2009), particularly in identifying alcohol and volatile acidity as key predictors of wine quality. While their models reported test RMSE values in the range of ~0.63–0.65, our tuned Random Forest and Gradient Boosting models achieved slightly better performance, with test RMSE values around 0.61–0.62 — despite using relatively simpler model architectures.

What sets our approach apart is the strong emphasis on hyperparameter tuning, feature selection, and robust validation. We incorporated techniques such as Out-of-Bag (OOB) scoring, Test MSE, and Test R^2 to evaluate generalization performance. This comprehensive and structured methodology reflects a more rigorous and professional modeling process than many standard baseline approaches.

In addition, while Cortez et al. built separate models for red and white wines, we combined them into a single dataset and found that the wine color variable had very little predictive importance when chemical properties were available. Overall, our modeling approach, which included careful feature scaling, hyperparameter tuning, and validation, produced competitive results and reinforced the conclusions drawn in prior research.

Conclusion:

In this project, we predicted wine quality using physicochemical features from combined red and white wine datasets. After preprocessing the data through scaling and encoding, we trained and evaluated both linear models (including Ridge and Lasso) and nonlinear models such as Decision Trees, Random Forests, and Gradient Boosting. Through hyperparameter tuning and feature selection analysis, we found that the tuned Random Forest model achieved the best performance. The results highlight that chemical properties like alcohol and acidity are strong predictors of wine quality, while wine color had minimal impact.

Reference

Cortez, Paulo, et al. “Modeling wine preferences by data mining from physicochemical properties.” *Decision Support Systems*, vol. 47, no. 4, Nov. 2009, pp. 547–553, https://doi.org/10.1016/j.dss.2009.05.016.

“Wine Quality.” *UCI Machine Learning Repository*, archive.ics.uci.edu/dataset/186/wine+quality. Accessed 7 May 2025.