CS 4120 – MACHINE LEARNING, DATA MINING

MIDPOINT REPORT

AUTHORS:

- **CHIZURUM EWELIKE**
- **LAYO FALOSEYI**

GITHUB: https://github.com/LAYO200/Wine-Quality-Prediction-models.git

WINE QUALITY MIDPOINT REPORT

Section 1.0: Overview

Expert tasters often judge wine quality, but their evaluations can be subjective and time-consuming. We build machine learning models to provide a reproducible, data-driven evaluation of wine quality, aiming to support consistent production decisions and offer reliable indicators for consumers.

Section 2.0: Dataset Description

We use the Wine Quality dataset from the UCI Machine Learning Repository (Cortez et al., 2009), containing 6,497 Portuguese Vinho Verde wine samples (1,599 red; 4,898 white). Each sample has 11 physicochemical features (e.g., acidity, residual sugar, pH, sulphates, alcohol) and a target quality score in the range of [0, 10] based on expert ratings.

Cleaning notes applied in code:

- Concatenated red and white subsets; added a categorical type column.
- Coerced all feature columns to numeric; dropped rows that became NaN after coercion.
- Cast quality to int; ensured features are float64 for stable MLflow schema.
- Derived binary label target_cls = 1 if quality ≥ 7, else 0.

Section 3.0: Tasks

- Classification: Predict High Quality (≥7) vs. Low Quality (<7).
- Regression: Predict the original quality score (0–10).

This dual-task design enables a side-by-side comparison of linear and nonlinear approaches on the same dataset.

Section 4.0: Metrics Plan

- Classification metrics: Accuracy and F1 score (to account for class imbalance). ROC-AUC may be reported as supplementary.
- Regression metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

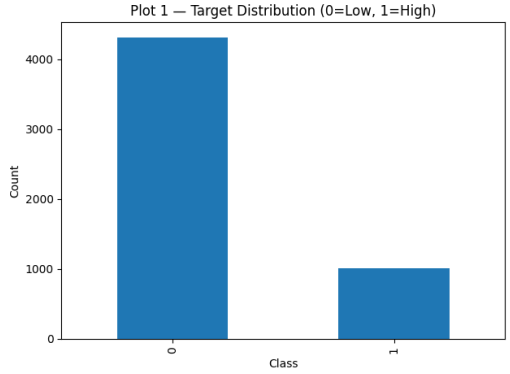
Model selection change: Our proposal planned a k-fold CV. For the midpoint, we switched to a fixed stratified train/validation/test split (seed = 42) to simplify iteration and runtime. We select the best classification model based on validation F1 and the best regression model based on validation RMSE, and we report the test metrics once. We will consider reinstating k-fold and small hyperparameter grids for the final.

Section 5.0: Baseline Plan (Classical ML Models)

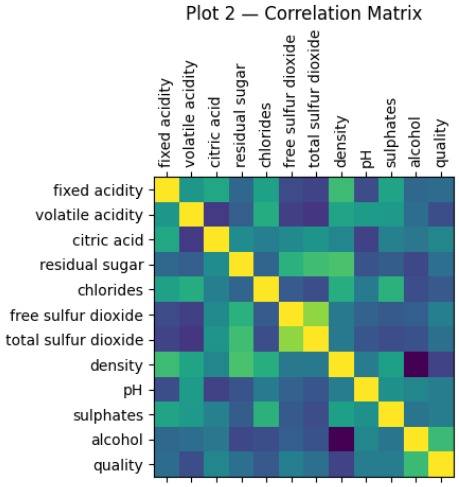
- Classification baselines: Logistic Regression (with standardization) and Decision Tree Classifier (no scaling).
- Regression baselines: Linear Regression (with standardization) and Decision Tree Regressor (no scaling).

Preprocessing is managed with scikit-learn Pipelines and a ColumnTransformer for scaling numeric features. We track parameters, metrics, figures, and serialized models in MLflow (with an input example for schema stability).

Section 6.0: Exploratory Data Analysis (EDA)



Interpretation: The distribution is imbalanced (≈ 81% class 0 vs 19% class 1). Due to this skew, accuracy alone can be misleading; therefore, we report F1 to reflect minority-class performance and use stratified splits.



Interpretation: Alcohol and sulphates show positive correlation with quality; volatile acidity tends to be negatively related. Several features are moderately correlated with each other (e.g., total sulphur dioxide and free sulphur dioxide), which motivates the use of regularized/linear baselines and tree-based models that handle interactions.

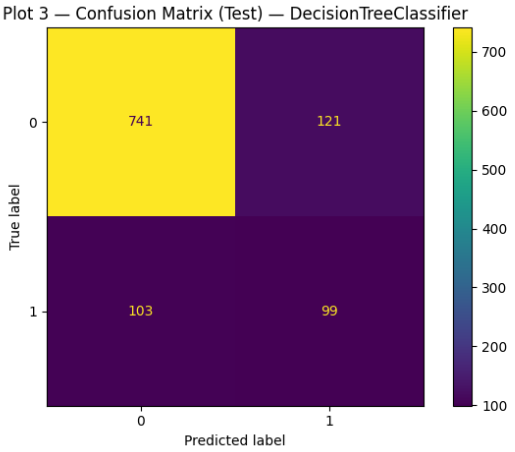
Section 7.0: Results

Table 1 — Classification metrics (Val Accuracy/F1; Test Accuracy/F1)

TABLE 1 – CLASSIFICATION METRICS (VAL/TEST)				
Model	Val_Accuracy	Val_F1	Test_Accuracy	Test_F1
Logistic Regression	*0.830827*	*0.430380*	*0.813910*	*0.369427*

Decision Tree Classifier	*0.793233*	*0.471154*	*0.789474*	*0.469194*

Interpretation: Decision Tree Classifier achieves a higher F1 (Val 0.471, Test 0.469) than Logistic Regression (Val 0.430, Test 0.369), so it's the best classifier by the primary metric under imbalance. Logistic regression has higher accuracy, but the performance of the minority class is weaker; hence, the preference for F1.

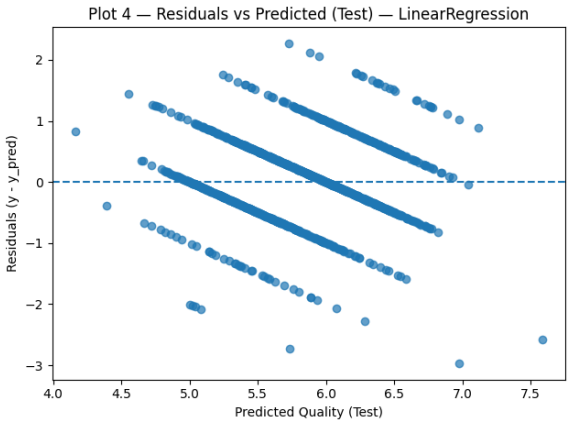


Interpretation: With TP=99, FN=103, FP=121, TN=741, errors cluster near the 6/7 threshold. False negatives (missed high-quality wines) are common when alcohol levels are moderate, but acidity/sulphates are favourable. False positives often have high alcohol but an unfavourable acidity profile—alcohol alone is not decisive. Precision and recall for class 1 are balanced (≈0.45–0.49), resulting in an F1 ≈ score of approximately 0.47.

Table 2 — Regression metrics (Val MAE/RMSE; Test MAE/RMSE)

Table 2 – Regression Metrics (Val/Test)				
Model	Val_Mae	Val_Rsme	Test_Mae	Test_Rmse
LinearRegression	0.573437	0.735946	0.546002	0.701434
Decision Tree Regressor	0.657895	0.959715	0.662594	0.960204

Interpretation: Linear Regression is best (Val MAE 0.573 / RMSE 0.736; Test MAE 0.546 / RMSE 0.701) vs. Decision Tree Regressor (Val 0.658 / 0.960; Test 0.663 / 0.960), indicating the target follows largely linear/monotonic trends that the linear model captures well.



Interpretation: Residuals exhibit a slight downward trend (under-prediction at low predicted scores and over-prediction at high), accompanied by mild heteroskedasticity (a wider spread at higher predicted values). Consider interaction terms or tree ensembles (and for the NN, nonlinearity) to reduce high-end bias.

Section 8.0: Results Summary:

- Classification: We select the Decision Tree as the top classifier, achieving an F1 score of 0.471 (Val) and 0.469 (Test), over Logistic Regression, which yields an F1 score of 0.430 (Val) and 0.369 (Test). Because the target is imbalanced, we prioritize F1 over accuracy (even though LR's accuracy is higher). The tree's win tells us we are benefiting from nonlinear interactions (e.g., alcohol × acidity × sulphates) that a linear boundary can't capture.
- Regression: We select Linear Regression as the best regressor (Val MAE 0.573 / RMSE 0.736; Test MAE 0.546 / RMSE 0.701), beating the Decision Tree Regressor (Val 0.658 / 0.960; Test 0.663 / 0.960). These results indicate largely monotonic/linear trends, and the linear model fits better.
- Class imbalance: We are working with a skewed label (≈ 81% class 0 vs 19% class 1). Accuracy can appear strong while minority performance lags, so we report F1 to reflect the quality of the positive class.
- Confusion matrix (best classifier on test): From Plot 3 (Decision Tree), we observe TP = 99, FN = 103, FP = 121, and TN = 741. Errors cluster at the 6/7 cutoff: we miss borderline 7s (false negatives, FN) and incorrectly promote some 6s (false positives, FP). This pattern highlights the limitations of shallow splits and single-feature cues (e.g., alcohol) in the absence of richer interactions.
- Residuals (best regressor on test): From Plot 4, we observe a slight downward trend in residuals (under-prediction at low predicted scores and over-prediction at high) and mild heteroskedasticity (a wider spread at higher predicted values).
- Next steps: For classification, we would try class weights, probability calibration, and tree ensembles (RF/GBDT). For regression, we would use linear regression as a strong baseline and explore interaction terms or an ensemble mitigator to reduce high-end bias.

Section 9.0: Neural Network Plan

We will implement a small MLP suitable for tabular data (e.g., 2 hidden layers with 32–64 units, ReLU activations), trained with the Adam optimizer and early stopping based on the validation loss. Inputs will be standardized. We will compare the NN against the classical baselines using the same splits and primary metrics (Acc/F1 for classification; MAE/RMSE for Regression).