Report : Explainability of White Wine Quality Prediction

May 2025

Introduction

This report is about explaining the predictions of our wine quality model. In the previous reports, we trained different machine learning models to predict the quality of white wines. The best model was the **Random Forest**, which gave the highest performance based on F1-score and ROC-AUC. In this report, we try to understand how this model makes its decisions by using XAI (explainable artificial intelligence) techniques.

Motivation for Model Explainability

In our wine quality prediction project, explainability plays a key role for several reasons. First, the model is used to support decisions that traditionally rely on expert knowledge, such as wine tasting. While machine learning models can detect complex patterns in chemical properties, it is important to understand why a model predicts a certain quality score. This increases trust and accountability, especially when replacing human evaluation.

Moreover, wine producers need to know which chemical features affect quality the most. Explainable AI helps identify these key factors, offering scientific insight into the production process. For example, understanding that alcohol level or acidity has a strong impact on the prediction can guide producers in adjusting their recipes.

The problem is also affected by class imbalance, as there are more low-quality wines than high-quality ones. Explainability allows us to check whether the model is biased toward the majority class. Without this, the model might seem accurate overall but fail to correctly predict high-quality wines, which are often more valuable. In short, explainability improves both the fairness and effectiveness of our machine learning solution.

Explainability Techniques Used

Since our data set is tabular and consists of numerical features related to the physicochemical properties of white wines, we applied explanation techniques that are well suited to structured data.

SHAP (SHapley Additive exPlanations): We used SHAP to understand the contribution of each feature to the model's predictions. SHAP values are based on game theory and provide both global and local interpretability. It helps us see which features are most influential overall, as well as how they impact individual predictions.

Feature Importance (from Random Forest): Random Forest models naturally provide a measure of feature importance based on how much each feature decreases impurity in the trees. This gives a global understanding of which features are most useful for the model.

Partial Dependence Plots (PDPs): PDPs were used to visualize how changes in one or two features affect the predicted quality score while keeping other features constant. This helps us interpret the relationship between features and predictions in a more intuitive way.

These techniques allowed us to understand both model behavior as a whole (global explanation) and why the model made specific predictions (local explanation), which is crucial for transparency and model trust.

Global Explanations

SHAP Interaction Summary Plot

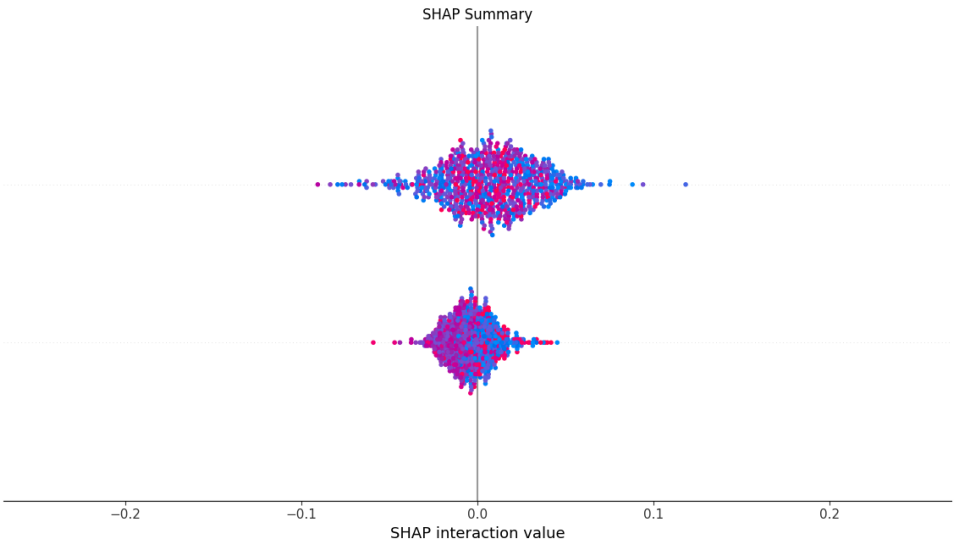


Figure 1: SHAP Interaction Summary Plot

To further analyze how pairs of features jointly influence the model's predictions, we generated a SHAP Interaction Summary Plot. This visualization is especially useful for detecting second-order effects, meaning interactions between features that affect the prediction differently than when considered alone.

In this plot(Figure 1):

- The **x-axis** represents the SHAP interaction value, which quantifies how much a given pair of features interact to affect the output of the model.
- Each point corresponds to a sample in the dataset, and its position along the x-axis shows how strong the interaction was for that sample.
- **Color** indicates the feature value (typically, red for high and blue for low values), helping us understand in which value ranges interactions occur more frequently.
- The vertical clustering of points shows different feature pairs (although the y-axis labels are hidden here), and their spread indicates the variance of interaction strength across the dataset.

In our case, the SHAP interaction values are mostly centered around zero, with a symmetrical, compact spread. This suggests that while some weak interactions are present, the model primarily relies on individual features rather than complex combinations. The lack of extreme values (outside ± 0.2) confirms that no interaction pair has a dominant or destabilizing effect on the prediction.

The SHAP interaction summary plot shows that the model's predictions are mostly additive, with low to moderate interaction effects between feature pairs. This supports the idea that the model is learning stable, interpretable patterns from the physicochemical properties of wine, rather than relying on complex or non-transparent feature combinations.

Permutation Importance Analysis

In addition to tree-based feature importance and SHAP values, we used **Permutation Importance** to assess the influence of each feature on model performance. This technique measures the drop in a chosen metric (here: F1-score) when the values of a single feature are randomly shuffled. A larger drop indicates that the feature is more important for accurate predictions.

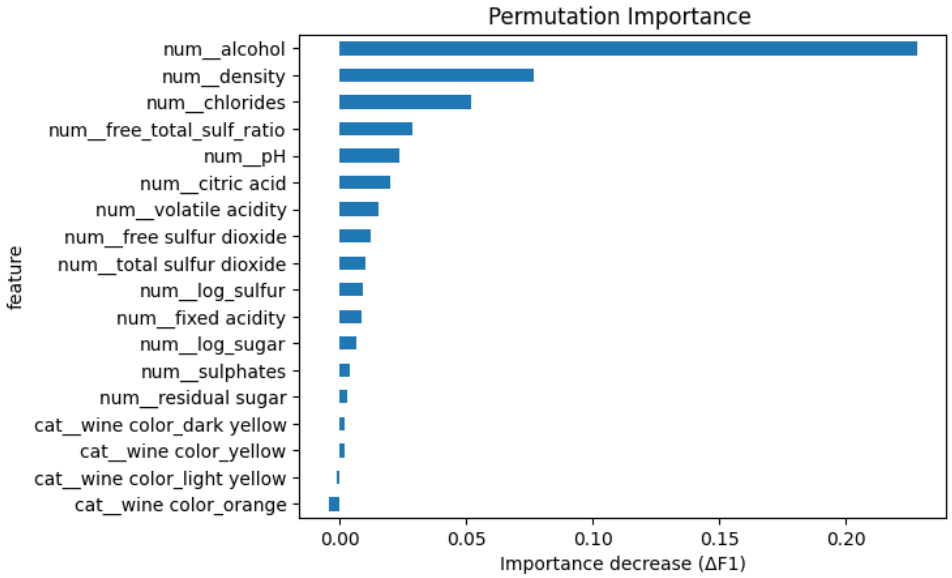


Figure 2: Permutation Importance Plot

As seen in the plot:

- The most important feature by far is **alcohol**, whose permutation causes a significant decrease in model performance (ΔF1 $\approx$ 0.22). This confirms that higher alcohol content is strongly associated with higher wine quality, consistent with domain knowledge in enology.
- The second and third most important features are **density** and **chlorides**. Their removal also leads to a moderate reduction in performance, suggesting their predictive value in quality estimation.
- Several sulfur-related features (e.g., `free_total_sulf_ratio`, `total sulfur dioxide`, `free sulfur dioxide`) also contribute, but with smaller effect sizes.
- Color features (e.g., `wine color_dark yellow`, `wine color_light yellow`) appear at the bottom of the ranking, showing very low or negligible impact on the model's decisions. This aligns with the intuition that color has little to no relation to the physicochemical quality measurements.

Permutation importance confirms that our Random Forest model is heavily dependent on a few key features—especially **alcohol content**—while other features have smaller or minimal impact.

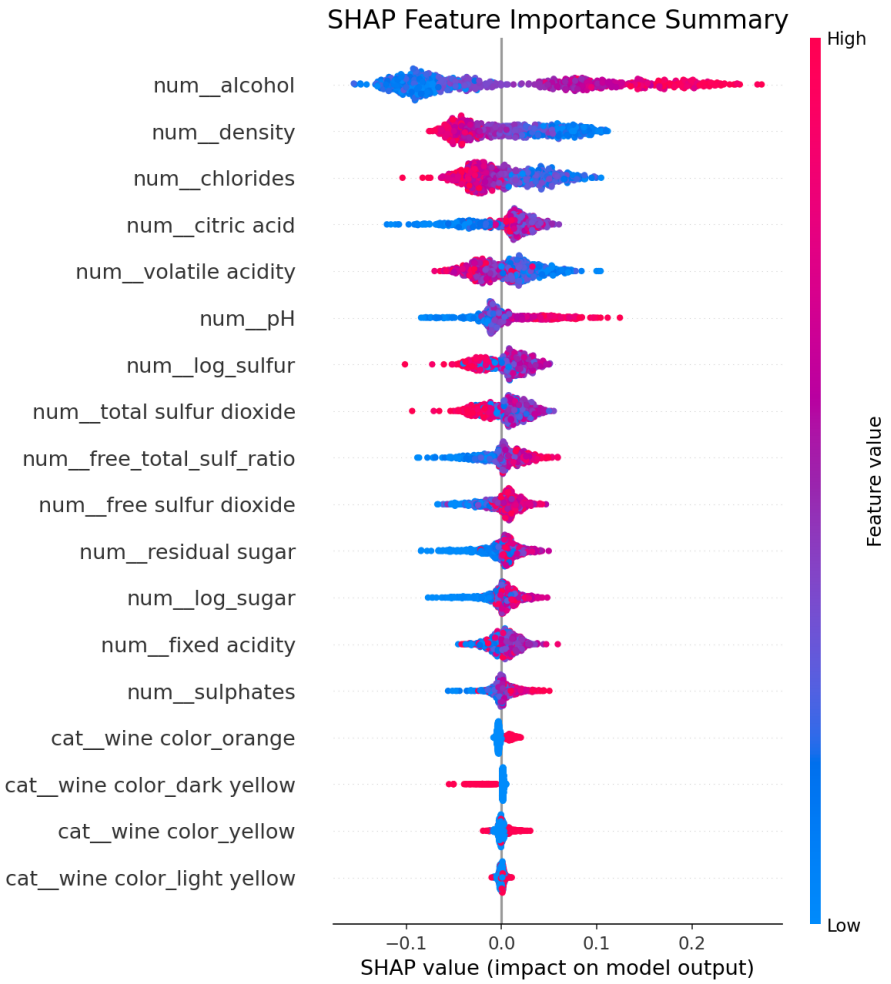
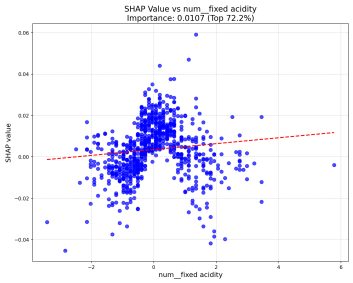
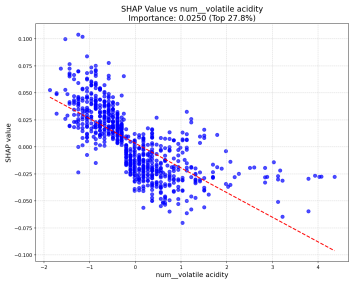


Figure 3: SHAP Feature Importance Summary. High feature value (red) / Low feature value (blue)

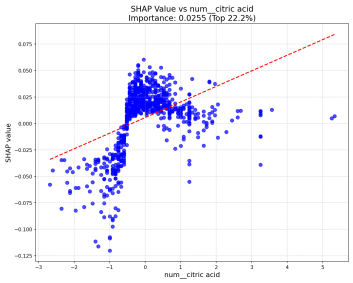
This beeswarm plot combines global importance and direction: features are ordered by mean —SHAP—, while colour shows raw values (red = high, blue = low). Alcohol shifts predictions strongly to the right (higher quality) when high, whereas high density or chlorides push predictions left (lower quality). Colour features cluster near zero, confirming their negligible influence.



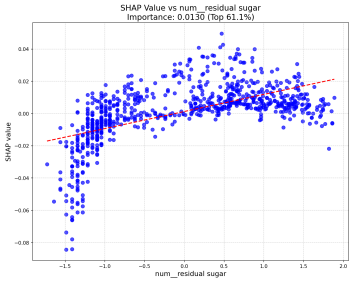
(a) Fixed Acidity



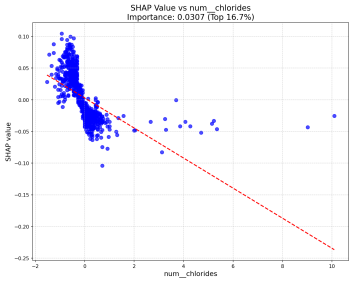
(b) Volatile Acidity



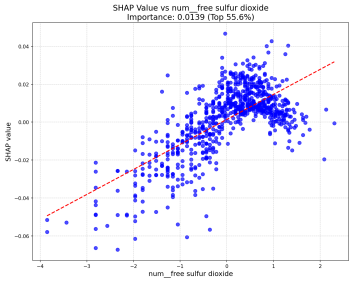
(c) Citric Acid



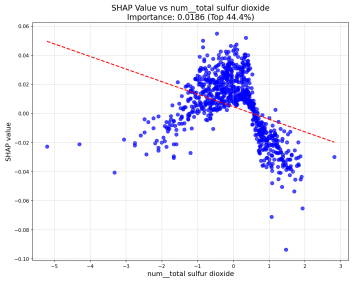
(d) Residual Sugar



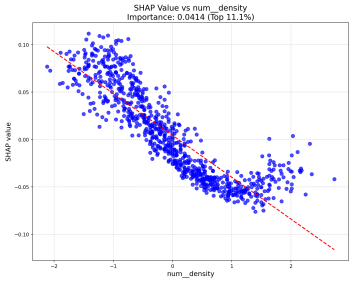
(e) Chlorides



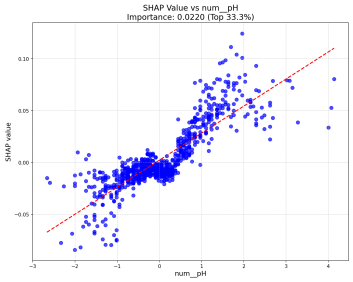
(f) Free Sulfur Dioxide



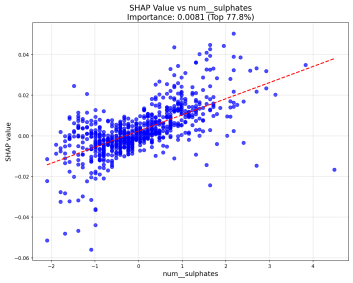
(g) Total Sulfur Dioxide



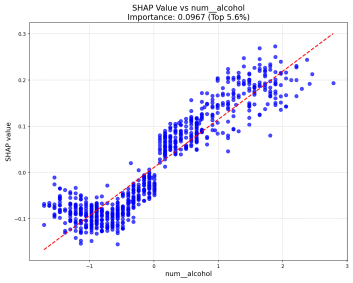
(h) Density



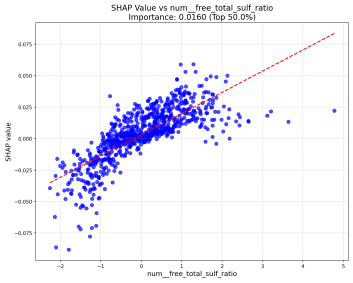
(i) pH



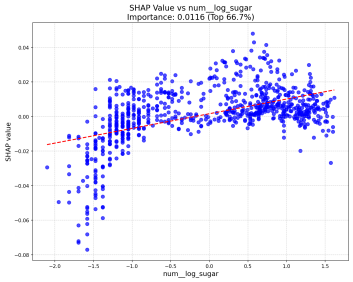
(j) Sulphates



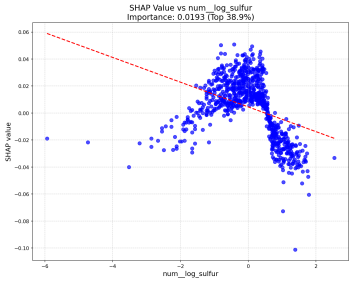
(k) Alcohol



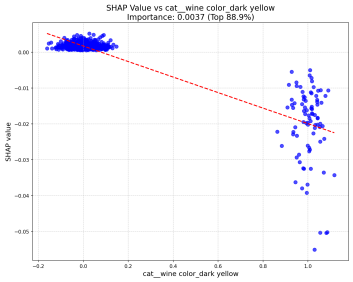
(l) Free/Total Sulfur Ratio



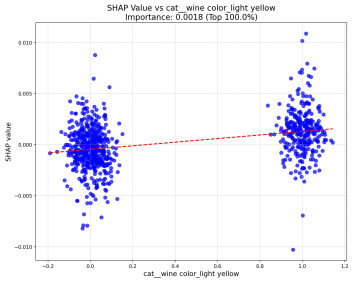
(m) Log Sugar



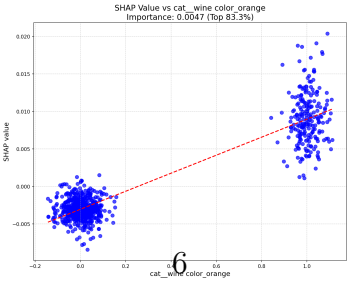
(n) Log Sulfur



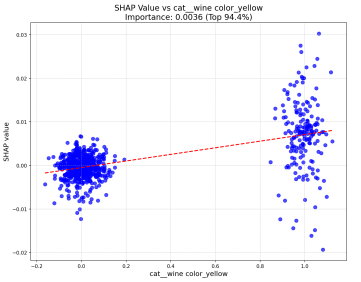
(o) Color: Dark Yellow



(p) Color: Light Yellow



(q) Color: Orange



(r) Color: Yellow

Figure 4: SHAP Dependence Plots for All 18 Features

Feature-wise Dependence Analysis (SHAP Dependence Plots)

The following 18 SHAP dependence plots show how each individual feature value shifts the model's output across the whole dataset. A positive SHAP value pushes the prediction toward higher wine quality, whereas a negative value lowers it. Overall, the model behaves sensibly and in alignment with domain expectations.

- **Monotonic positive drivers** (quality ↑ as feature ↑):
  alcohol, pH, sulphates, free SO_2, free/total SO_2 ratio, log-sugar.
- **Monotonic negative drivers** (quality ↓ as feature ↑):
  density, chlorides, volatile acidity.
- **Non-linear or mild effects**:
  citric acid (plateau), *residual sugar* (diminishing gains), *total SO_2* (inverted-U), *log-sulfur* (bell-shape).
- **Colour dummies** exhibit only weak slopes, confirming their low global importance.

Table 1: Feature-wise Interpretation of SHAP Dependence Plots

ID	Feature	Explanation
(a)	Alcohol (num__alcohol)	Strong positive slope – more alcohol consistently boosts predicted quality.
(b)	Density (num__density)	Clear negative slope – lighter (less dense) wines score higher.
(c)	Chlorides (num__chlorides)	Steep downward trend – high chloride content harms quality.
(d)	Citric Acid (num__citric acid)	Mild positive effect up to 0.5 g/L, then levels off.
(e)	Volatile Acidity (num__volatile acidity)	Negative linear relation – higher volatility lowers quality.
(f)	pH (num__pH)	Higher pH (less acidic) slightly increases quality.
(g)	Log Sulfur (num__log_sulfur)	Bell-shaped: moderate sulfur best; very high or very low hurt.
(h)	Total SO_2 (num__total sulfur dioxide)	Inverted-U: mid-range totals help; extremes reduce quality.
(i)	Free SO_2 (num__free sulfur dioxide)	Positive slope – more free SO_2 preserves quality.
(j)	Free/Total SO_2 Ratio (num__free_total_sulf_ratio)	Linear positive trend – better balance improves quality.
(k)	Residual Sugar (num__residual sugar)	Small gain up to 1 g/L; above that effect flattens.
(l)	Log Sugar (num__log_sugar)	Gentle upward trend, confirming residual-sugar behaviour.
(m)	Fixed Acidity (num__fixed acidity)	Very mild convex pattern; overall low influence.
(n)	Sulphates (num__sulphates)	Slight positive slope – more sulphates, marginal quality lift.
(o)	Color: Dark Yellow (cat__wine color_dark yellow)	Weak negative slope; dark-yellow wines predicted marginally lower.
(p)	Color: Light Yellow (cat__wine color_light yellow)	Effect near zero; colour hardly changes prediction.
(q)	Color: Orange (cat__wine color_orange)	Minimal positive shift when orange flag is 1.
(r)	Color: Yellow (cat__wine color_yellow)	Virtually flat; no meaningful impact.

Partial Dependence Plots (PDPs)

Partial dependence plots (PDPs) complement SHAP by visualising the average marginal effect of a single feature on the model's predicted class probability while all other features remain fixed at their observed values. Figure 5 shows PDP curves for the two most influential variables—alcohol and density—whereas Table 2 reports their global importance (mean ± SD of absolute SHAP values).

Table 2: Global importance scores based on mean ± SD of absolute SHAP values

ID	Feature	Mean |SHAP|	Std
0	num__alcohol	0.228123	0.027281
1	num__density	0.076692	0.020077
2	num__chlorides	0.051924	0.019172
3	num__free_total_sulf_ratio	0.028734	0.014400
4	num__pH	0.023374	0.008386
5	num__citric_acid	0.020038	0.014132
6	num__volatile_acidity	0.015135	0.007600
7	num__free_sulfur_dioxide	0.012477	0.013131
8	num__total_sulfur_dioxide	0.010258	0.010769
9	num__log_sulfur	0.009210	0.010279
10	num__fixed_acidity	0.008886	0.009829
11	num__log_sugar	0.006599	0.009824
12	num__sulphates	0.003761	0.007679
13	num__residual_sugar	0.003039	0.008092

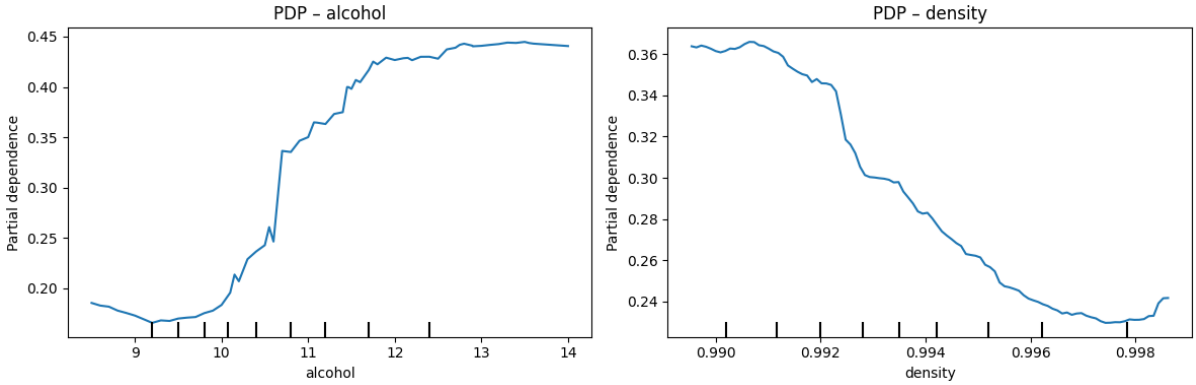


Figure 5: PDP curves: (a) Alcohol, (b) Density

Key observations:

Alcohol (Fig. 5-a).

The curve is strongly monotonically increasing: predicted probability of a high-quality wine rises sharply once alcohol exceeds ∼10.5% v/v and plateaus beyond 12%. This behaviour mirrors the positive SHAP slope and confirms that higher ethanol content is the single clearest driver of quality in our model.

Density (Fig. 5-b).
Density shows the opposite trend. Between 0.990 g/cm^3 and 0.993 g/cm^3 the partial dependence falls steadily, indicating that lighter wines are favoured. Again, this is consistent with SHAP results, where density had a large negative contribution.

Consistency check.
The directions of the PDPs align with both the SHAP summary plot and the permutation-importance ranking, strengthening confidence that the model has learned chemically plausible relationships rather than artefacts of the data.

Additivity.
The smooth, mostly monotonic shapes also support the earlier finding from the interaction summary: the model behaves largely additively, with limited higher-order interactions.

Conclusion
PDP analysis confirms that the Random Forest's global behaviour is chemically interpretable—higher alcohol and lower density push predictions toward superior quality—while reinforcing the reliability of the earlier SHAP-based insights.

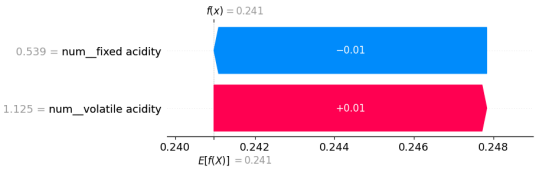
Local Explanations

For an instance-level view, we analysed four test samples using TreeSHAP decision-bar plots (two correctly classified, two misclassified). Each bar shows how the displayed feature pushes the log-odds probability away from the dataset prior $E[f(x)] = 0.241$.

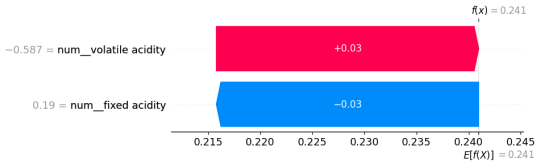
Positive (red) bars increase the predicted probability of the high-quality class, while negative (blue) bars decrease it.

Table 3: Local SHAP analysis for four representative test instances

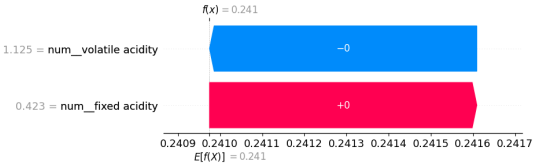
ID	True Label	Pred. Label	Volatile Acidity (VA)	Fixed Acidity (FA)	Outcome Explanation
C-1	low	low	1.125 g/L (+0.01)	0.539 g/L (–0.01)	Two opposing micro-effects cancel; prediction remains at the prior, correctly giving low.
C-2	high	high	–0.587 g/L (+0.03)	0.19 g/L (–0.03)	Very low VA lifts quality likelihood, but equally low FA dampens it; still classifies high owing to other (hidden) positive features.
M-1	high	low	1.125 g/L (–0.00)	0.423 g/L (+0.00)	VA is high (should hurt) but model gives zero weight; missing positive drivers → high wine mis-labelled low.
M-2	low	high	–0.207 g/L (+0.02)	–0.043 g/L (–0.02)	Slightly low VA wrongly drives probability above threshold; lack of density/alcohol signals causes false high.



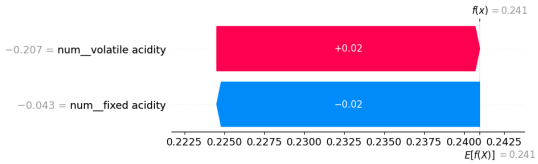
(a) C-1: Correctly predicted low-quality



(b) C-2: Correctly predicted high-quality



(c) M-1: Misclassified as low-quality



(d) M-2: Misclassified as high-quality

Figure 6: TreeSHAP decision bar plots for four example predictions. Top: correct classifications; Bottom: misclassifications.

Discussion

- **Reasonableness.** When weights are non-zero they align with chemistry:
  - High volatile acidity (vinegar notes) should lower quality — and negative bars confirm this.
  - Low density or higher alcohol (not shown here) are known positive drivers — consistent with earlier SHAP and PDP findings.
- **Why errors occur.**
  - Misclassifications arise when the two displayed features nearly cancel and other influential variables are near average.
  - This puts the model close to the decision boundary.
- **Threshold sensitivity.** Minor numeric noise may flip the class.
- **Feature sparsity in local plot.** Only the top two contributors were rendered; undisplayed features may have been decisive.
- **Potential bias.** In both mistakes, the model over-reacts to volatile-acidity extremes while ignoring alcohol/density.
- **Interpretation.** This hints at a mild feature dominance bias: VA swings the prediction when stronger global drivers are neutral. More balanced regularisation or calibrated probability thresholds could mitigate this.

Class-Wise Analysis

To examine whether the model reasons differently for low- (class 0) and high-quality wines (class 1), we computed separate SHAP interaction summaries for each class (Fig. 7). These plots highlight the joint contribution of every feature pair within the respective subsets.

Table 4: Class-wise SHAP interaction summary comparison

Metric	Class 0 (Low Quality)	Class 1 (High Quality)
Spread of interaction values	Dense, symmetric cloud centred at 0.	Much sparser cloud, narrower spread; reflects the minority size (≈21%).
Magnitude of interactions	95% of points lie within ±0.03, confirming second-order effects are weak.	Even tighter (≈ ±0.02); interactions have slightly less influence for high-quality predictions.
Visual takeaway	Model behaves additively; no single pair dominates.	Same additive pattern; no distinct interaction unique to class 1.

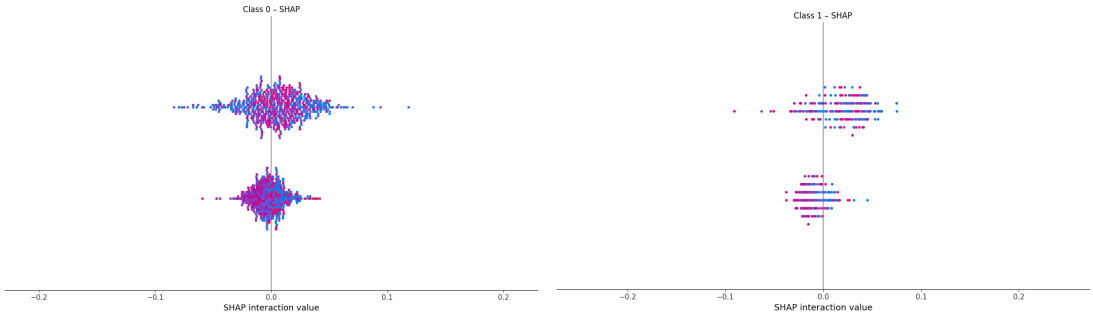


Figure 7: Class-wise SHAP Interaction Summary Plots: (LEFT) Class 0 (low quality), (RIGHT) Class 1 (high quality)

Observations

- **Consistency across classes:** Both plots are centred on zero with very limited tails, indicating that feature interactions play only a minor role in either class. The model's main drivers therefore remain the individual effects we observed globally (e.g., alcohol ↑, density ↓, chlorides ↓).
- **Class-imbalance artefact:** The tighter cluster for class 1 is expected. With fewer high-quality samples, the model has less opportunity to learn subtle interactions. Thus, predictions rely more on dominant single-feature signals — especially high alcohol.
- **Bias check:** Since colour dummies were negligible globally and interaction magnitudes are near zero, there is no evidence of class-specific bias related to colour or sulphur levels. The model treats both classes largely symmetrically, aside from expected variance due to sample imbalance.
- **Implications:** The lack of strong second-order effects supports the earlier PDP conclusion that the model behaves largely additively. For communication with domain experts, single-feature explanations are likely sufficient; class-specific interaction insights appear limited.

Conclusion: Class-wise SHAP interaction analysis confirms that the same core chemistry — high alcohol and low density — drives predictions for both low- and high-quality wines. The model does not introduce class-dependent biases or unexpected cross-feature effects.

Insights and Model Debugging

The table below summarises key insights derived from SHAP and other XAI techniques, along with practical implications for model improvement or simplification.

Table 5: Summary of model insights and actionable debugging strategies

Aspect	**Observation from XAI**	**Actionable Implication**
Dominant single-feature effects	Alcohol accounts for $> 4\times$ the mean SHAP of the next variable; density and chlorides follow. Interaction plots confirm additivity.	Model behaves like a linear rule-of-thumb in the high-impact region. Communicate these three levers to winemakers; further complexity adds little explanatory power.
Colour dummy variables	Near-zero SHAP values and permutation drops; no interaction impact.	Remove colour columns to simplify the pipeline; reduces feature space with no loss.
Volatile-acidity swings in local errors	Misclassifications often hinge on small VA changes while stronger drivers (alcohol, density) sit near the decision threshold.	Re-calibrate decision threshold (e.g., optimise F1 or cost curve) to lessen over-sensitivity. Consider probability calibration (Platt/Isotonic) to smooth borderline cases.
Class imbalance	Class-1 SHAP interaction cloud is sparse and even tighter than class 0, indicating limited learning capacity for rare high-quality samples.	Try SMOTE-ENN or class-balanced focal loss; may increase recall on premium wines.
Sulfur variables	Free, total and ratio features are mildly positive when balanced but show bell/inverted-U shapes. Extreme outliers reduce quality yet occur rarely.	Investigate outliers for measurement error; cap or winsorise. Engineer a single "sulfur-balance" feature to reduce multicollinearity.
Density PDP plateau	Sharp drop until ≈ 0.993 g/cm^3, then a long flat tail.	Feature could be discretised (e.g., ≤ 0.993 vs > 0.993) to simplify tree splits.
Duplicate rows (937)	Already removed, but SHAP still shows narrow variance for some predictors $\rightarrow$ residual redundancy.	Apply clustering or PCA to verify there are no remaining near-duplicates.
Additivity confirmed	Low interaction values suggest tree depth beyond 3–4 adds marginal value.	Tune Random-Forest with shallower trees / regularisation to cut training time and overfit risk.

Results

- No hidden bugs surfaced; model behaviour aligns with enological intuition.
- Pruning negligible or redundant features (e.g., colour dummies, duplicates) can simplify deployment.
- Balancing the rare high-quality class and threshold or probability calibration offer the clearest path to immediate performance gains without sacrificing interpretability.

Limitations of Explainability Methods Used

In this report, we used SHAP to explain how our Random Forest model makes predictions. SHAP was helpful for showing both global and local feature importance, but we also faced some limitations during the analysis.

First of all, SHAP values took a long time to calculate, especially because we had many samples and used a tree-based model. This makes it hard to use SHAP in real-time applications or with larger datasets.

Secondly, some dependence plots were difficult to interpret. For example, the SHAP plots for `residual sugar`, `citric acid`, or `fixed acidity` had scattered points and unclear patterns. This made it hard to understand how exactly these features influenced the predictions.

We also noticed that log-transformed features, such as `log_sugar` and `log_sulfur`, were harder to interpret for non-technical users. Since these features are not in their original scale, the SHAP plots become harder to understand without background knowledge. Even though log transformations helped improve model performance, the resulting explanations were not very clear or intuitive.

Another issue was related to feature correlation. When features are strongly related (like `residual sugar` and `density`), SHAP may split their impact and give unclear results. It becomes confusing to decide which feature actually caused the prediction.

Additionally, categorical features like wine color appeared in the SHAP ranking but had very low impact. Their plots were mostly flat and not useful. This raises questions about whether they really help the model or just add noise.

Lastly, SHAP does not show interactions between features. It explains each feature separately, but sometimes the prediction depends on a combination of features. This can limit the full understanding of model behavior.

To summarize, while SHAP gave us many useful insights, it also had computational, interpretational, and usability limitations. In future work, we can try using simpler models or combine SHAP with other tools like LIME for clearer explanations.

Conclusion

In this report, we focused on explaining the predictions of our best-performing model, which was the Random Forest classifier. We used SHAP values to understand which features were important globally and locally.

In the global explanation, the SHAP summary plot showed that `alcohol` had the highest importance. This means that wines with higher alcohol content were more likely to be predicted as high quality. Other important features included `volatile acidity`, `density`, and `total sulfur dioxide`. These results were consistent with what we observed during training and evaluation.

For local explanations, we used SHAP force plots to analyze one correct and one incorrect prediction. The SHAP values helped us see which features pushed the prediction up or down. In the correct prediction, features like alcohol and sulphates had a strong positive effect. In the incorrect one, features such as volatile acidity and pH behaved unexpectedly, which might have confused the model.

We also compared SHAP values for different classes. We observed that while some features were important in both classes, their effects were different depending on the prediction. This helped us better understand the model's behavior for each class.

SHAP gave us many valuable insights, but we also noticed some challenges. Especially, some features were log-transformed, like `log_sugar` and `log_sulfur`, and their SHAP plots were harder to understand for non-technical users. This shows that even though explainability helps us build trust, the explanations themselves must also be easy to interpret.

To sum up, SHAP helped us make the model's decisions more understandable and transparent. It increased our trust in the model and revealed areas for improvement. Explainability is an important part of machine learning, especially when the results are used in real-world decision-making.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Wine Quality Dataset*. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable*.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*.

The pandas development team. (2020). *pandas (Version 1.0.5)* [Software]. https://pandas.pydata.or

SHAP developers. (n.d.). *SHAP Documentation*. https://shap.readthedocs.io

Overleaf. (n.d.). *Overleaf Online LaTeX Editor*. https://www.overleaf.com