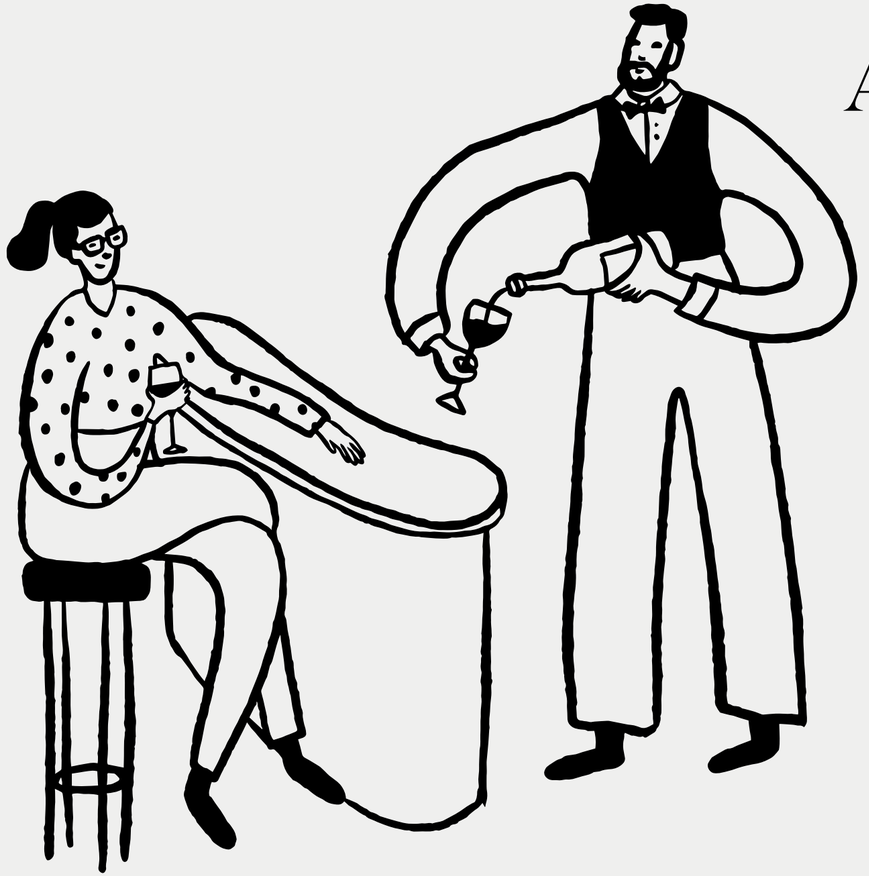
Key Factors of Wine Quality:

A Marketing Strategy Case Study for Vivino



Lopa Detroja
Kitsana Sudsaneh

Table of Contents

Introduction and Executive Summary

Goals

- Identify key physicochemical properties from Vivino's dataset of 1,599 French Bordeaux wines to classify wine quality.
- Use machine learning models to predict and understand the factors defining a "good" wine.

Methods

- Used **Naive Bayes** (a probabilistic model) and **Random Forest** (an ensemble decision tree method) for classification.

Key Findings

- **Random Forest** outperformed Naive Bayes in accuracy and feature analysis.
- **Top 3 features:**
  - **Alcohol:** Correlated with higher quality.
  - **Sulphates:** Enhanced flavor and preservation.
  - **Volatile Acidity:** Lower levels indicate better quality.

Data Preparation:
(Cleaning and Preparing)

Data Overviews

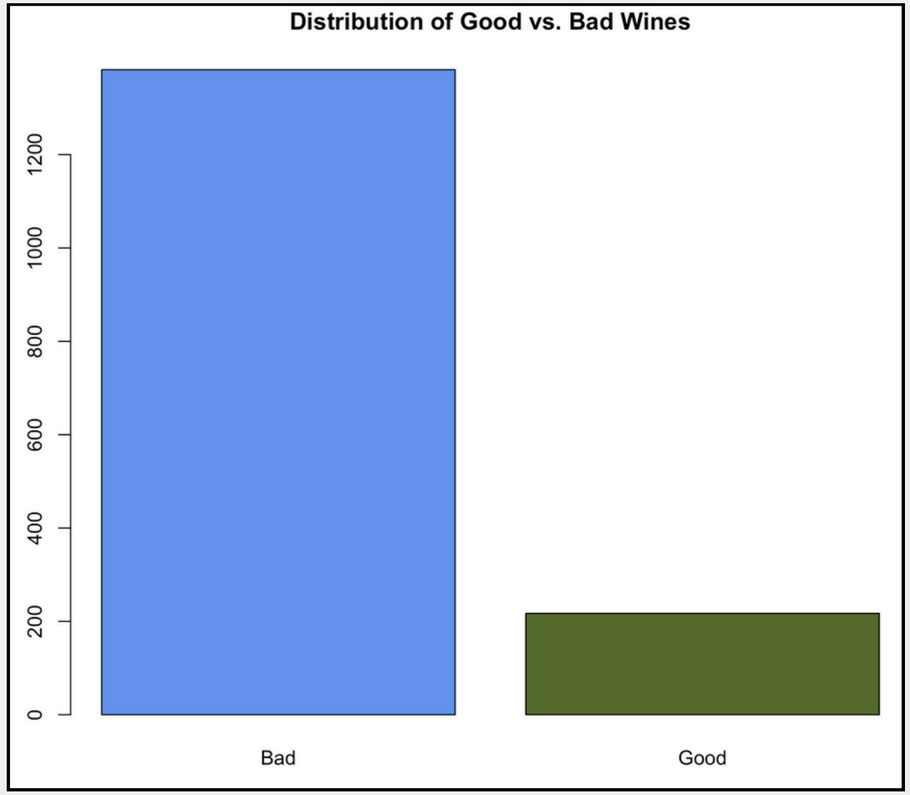
- 1,599 observations of French Bordeaux wines.

- 11 physicochemical properties
  - fixed acidity
  - volatile acidity
  - citric acid
  - residual sugar
  - chlorides
  - free sulfur dioxide
  - total sulfur dioxide
  - density
  - pH
  - sulphates
  - alcohol
- and one quality score.

Data Preparation

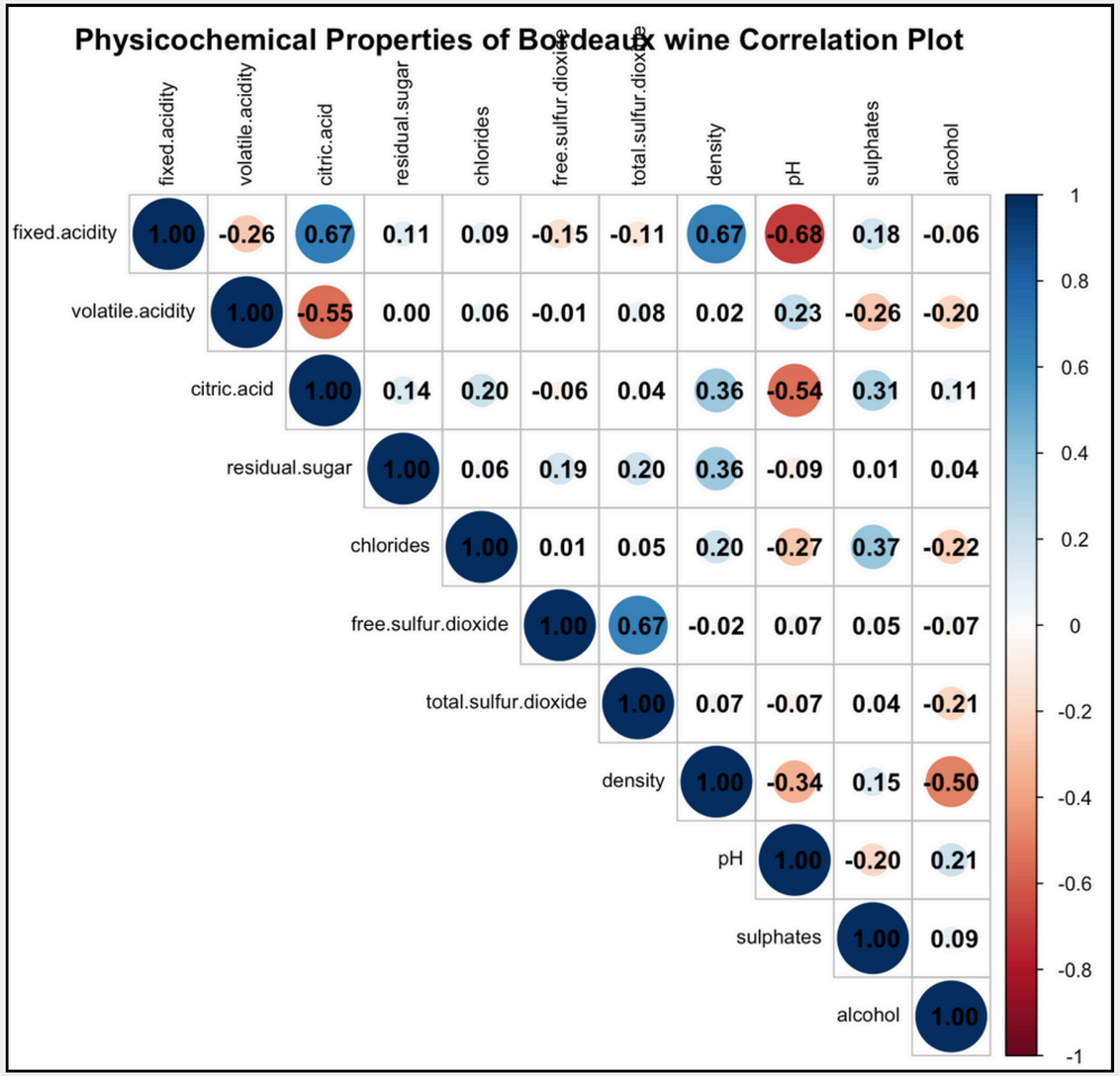
- **Dataset Loaded:**
  - 1,599 rows and 12 columns.

- **Target Variable Created:**
  - "Good" wine: Quality > 6 (217 samples).
  - "Bad" wine: Quality ≤ 6 (1,382 samples).



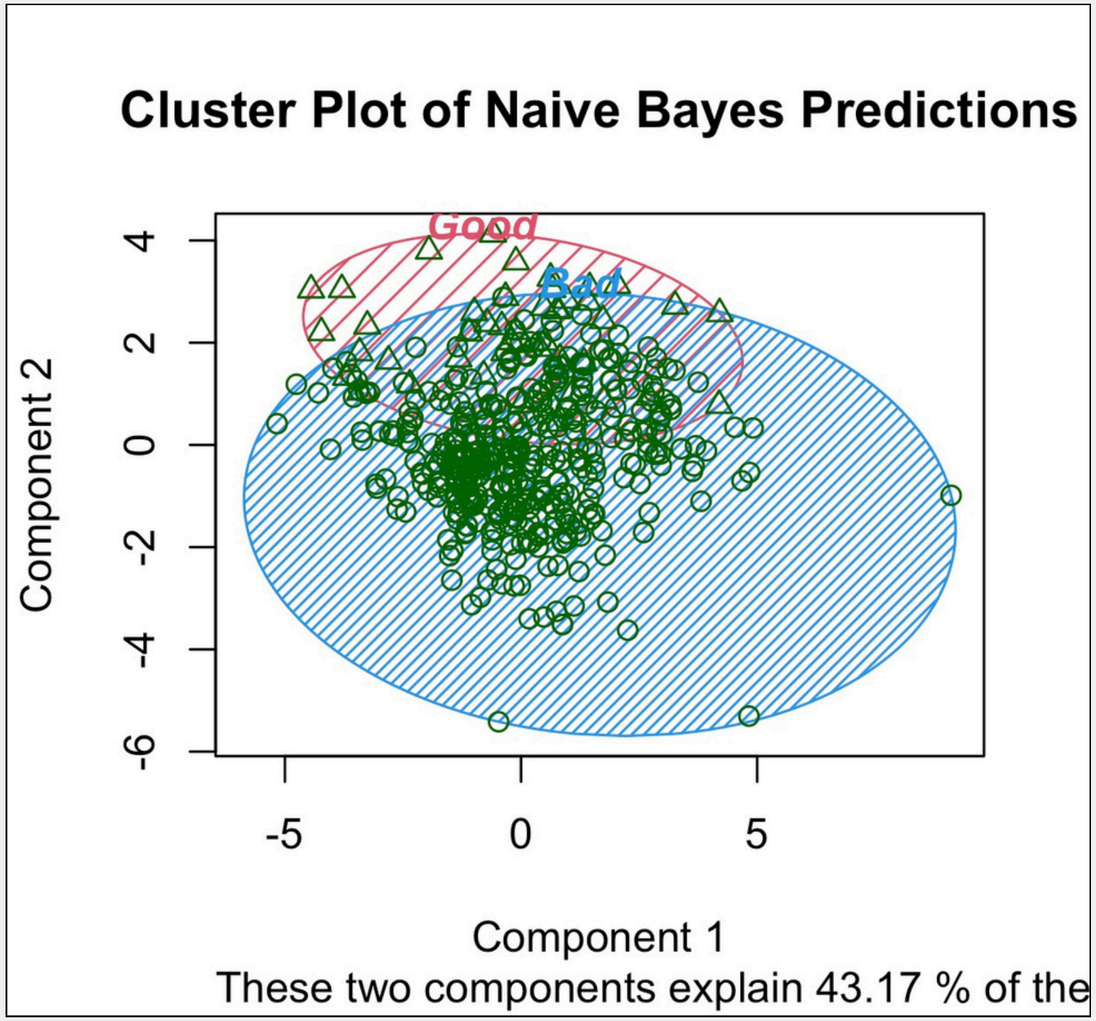
- **Train-Test Split:**
  - Training set (70%): 1,119 samples (956 "Bad," 163 "Good").
  - Testing set (30%): 480 samples (426 "Bad," 54 "Good").

Exploratory Data Analysis



- **Positive Correlation:**
- A strong positive correlation indicates that as the amount of free sulfur dioxide increases, the total sulfur dioxide also increases proportionally.
- Example: Total Sulfur Dioxide and Free Sulfur Dioxide (r = 0.67)
- **Negative Correlation:**
- A strong negative correlation indicates that as the acidity of the wine increases (lower pH), the pH value decreases. This is expected because pH measures acidity inversely.
- Example: pH and Fixed Acidity (r = –0.68)

Naive Bayes Model



Training Model Details

- Dataset: The dataset consists of 1,599 observations of French Bordeaux wines with 11 variables. We focus on predicting wine quality, where the target variable quality.label is categorized as "Good" (quality > 6) or "Bad" (quality ≤ 6).
- **Model Type:** Naive Bayes (a probabilistic classifier).
- **Split Dataset:** Training Dataset: 70% of the data (1,119 samples), Testing Dataset: 30% of the data (480 samples).
- We applied a Naive Bayes classifier, a probabilistic model based on Bayes' Theorem, to predict whether a wine is "Good" or "Bad."

Naive Bayes Model

Training Model Details

- Built a Naive Bayes model for classification to predict wine quality.
- Split the dataset: 70% training data (1,119 samples) and 30% testing data (480 samples).

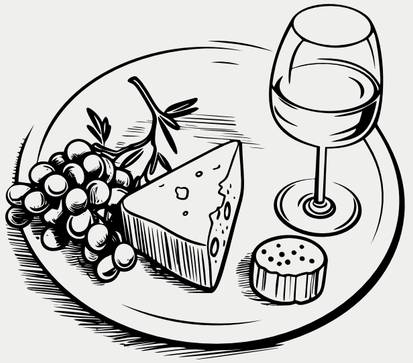
Before Handling Imbalance (Original Model)

```
Confusion Matrix:
> conf_matrix <- table(wine.nb.pred, wine.df.test$quality.label)
> print(conf_matrix)

wine.nb.pred Bad Good
        Bad  359   25
        Good  67   29
```

- **Class Error:** "Bad" wines had low error (63.21%), while "Good" wines had a higher error (36.79%).
- **Model Bias:** The model had a slight bias towards the majority class ("Bad").
- **Cluster Plot:** Significant overlap between "Good" and "Bad" classes, reflecting poor separation for the "Good" wines.

Accuracy: 80.83%



Naive Bayes Model

Issue with Imbalanced Classes

- The dataset is imbalanced with more "Bad" wines than "Good" wines, affecting prediction accuracy for the "Good" wines.

After Handling Imbalance (Balanced Sampling):

- **Balanced Sampling:**
  - ensures the model equally focuses on both classes, improving the accuracy for the "Good" class.
- **Class Error:**
  - Slightly reduced error for both classes.
- **Adjusted Rand Index (ARI):**
  - ARI: 0.21, indicating a modest agreement between the predicted and actual classes.
- **Cluster Plot:**
  - Improved separation between the "Good" and "Bad" classes after balancing the dataset.

Naive Bayes Model

Performance Metrics

	RESULT	EXPLANATION
Accuracy	80.83%	• Naive Bayes provides a solid prediction performance with an accuracy of 80.83%, while Random Forest slightly outperforms it at 82.9%.
Adjust Rand Index	0.21	• The Adjusted Rand Index (ARI) indicates a moderate agreement between predicted and actual labels, though it suggests room for improvement, especially for predicting the "Good" class.
Confusion Matrix	Error rate: 20.37% (Good)	• Among the "Bad" wines in the test data, Naive Bayes correctly classified 359 "Bad" wines and misclassified 67 as "Good". Among the "Good" wines, 29 were classified as "Bad".

Top Features: Naive Bayes identifies Alcohol, Residual Sugar, and Sulphates as the most important features in predicting wine quality.

Naive Bayes Model

Summary Function

```
> print(summary_pred)
  Group.1 fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides free.
sulfur.dioxide
1     Bad      7.929948        0.5713932   0.2259115       2.517187 0.09032552
17.08203
2    Good      9.607292        0.3648958   0.4489583       2.556250 0.07645833
12.38542
  total.sulfur.dioxide   density       pH sulphates  alcohol  quality quality.label
1             51.92057 0.9966970 3.333880 0.6425260 10.20234 5.502604      1.065104
2             33.17708 0.9965201 3.244687 0.7305208 11.53611 6.208333      1.302083
```

```
> print(summary_actual)
  Group.1 fixed.acidity volatile.acidity citric.acid residual.sugar  chlorides free.
sulfur.dioxide
1     Bad      8.261737        0.5419953   0.2649765       2.498944 0.08976761
15.97770
2    Good      8.294444        0.4362037   0.3142593       2.730556 0.07007407
17.44444
  total.sulfur.dioxide  density       pH sulphates  alcohol  quality quality.label
1             48.73826 0.996787 3.312230 0.6499296 10.32113 5.464789             1
2             43.70370 0.995672 3.346111 0.7405556 11.63642 7.055556             2
>
>
```

3 Tops properties

- **Alcohol:**
  - Predicted: 11.54 for "Good" wines vs. 10.20 for "Bad" wines.
  - Actual: 11.63 for "Good" wines vs. 10.32 for "Bad" wines.
  - Observation: Higher alcohol content is strongly associated with "Good" wines.
- **Sulphates:**
  - Predicted: 0.73 for "Good" wines vs. 0.64 for "Bad" wines.
  - Actual: 0.74 for "Good" wines vs. 0.65 for "Bad" wines.
  - Observation: "Good" wines consistently have higher sulphate levels, which enhance flavor and preservation.
- **Volatile Acidity:**
  - Predicted: 0.36 for "Good" wines vs. 0.57 for "Bad" wines.
  - Actual: 0.43 for "Good" wines vs. 0.54 for "Bad" wines.
  - Observation: Lower volatile acidity improves the taste, making wines more likely to be classified as "Good."

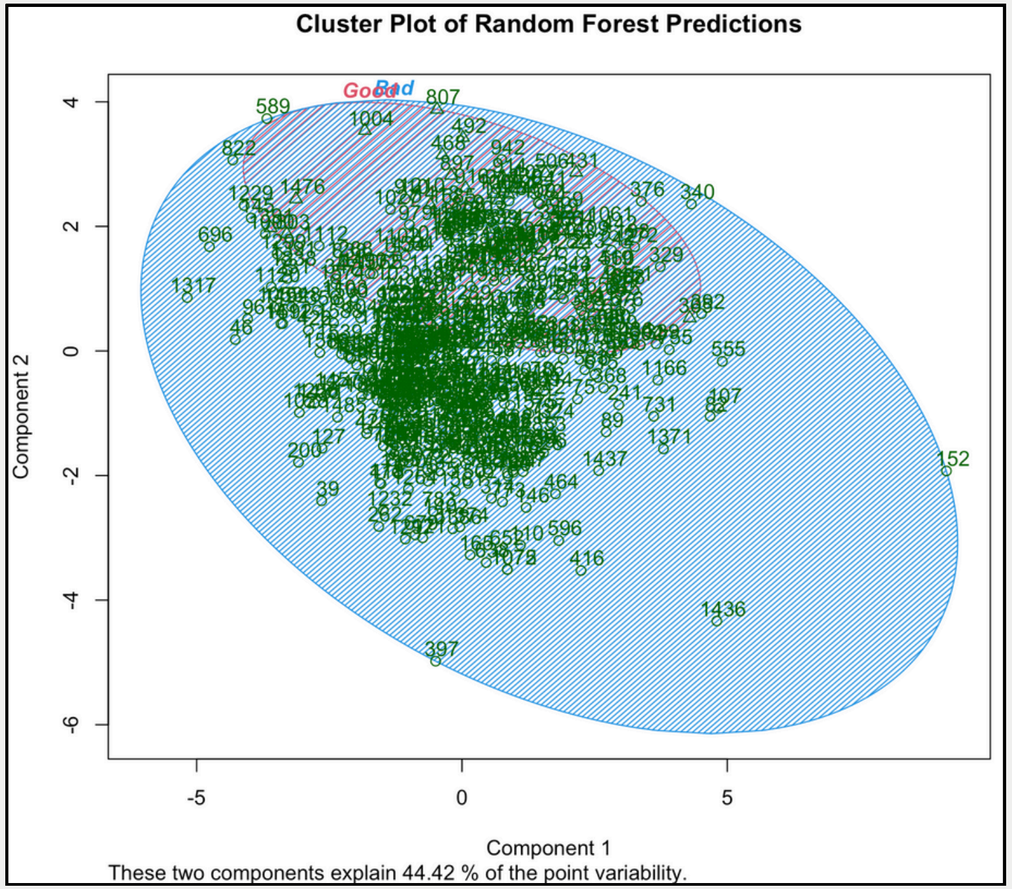
Random Forest Model

Training Model Details

- Built a Random Forest model with 3,000 decision trees for classification.
- Split dataset: 70% training data (1,119 samples) and 30% testing data (480 samples).

Before Handling Imbalance (Original Model)

- **Class Error:** "Bad" wines had low error (2.65%), while "Good" wines had high error (46.01%).
- **Model Bias:** The model performed significantly better for the majority class ("Bad") compared to the minority class ("Good").
- **Cluster Plot:** Significant overlap between "Good" and "Bad" clusters, highlighting poor separation for "Good" wines.




```
        OOB estimate of  error rate: 8.94%
Confusion matrix:
     Bad Good class.error
Bad  931   25  0.02615063
Good  75   88  0.46012270
```

Random Forest Model

Issue with Imbalanced Classes

- The dataset is imbalanced, with more "Bad" wines (1,382) than "Good" wines (217). The model prioritizes "Bad" wines, leading to poor performance on "Good" wines.

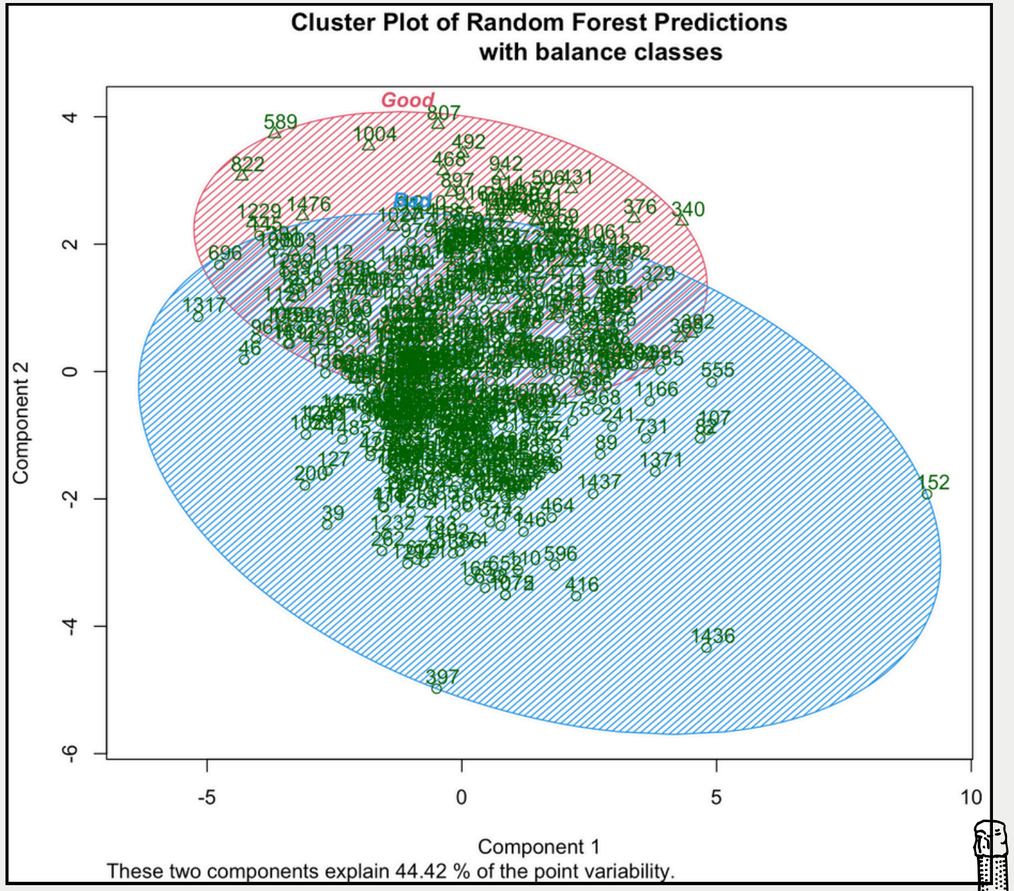
- Balanced sampling ensures the model focuses equally on "Good" wines, improving their classification accuracy.

After Handling Imbalance (Balanced Sampling):

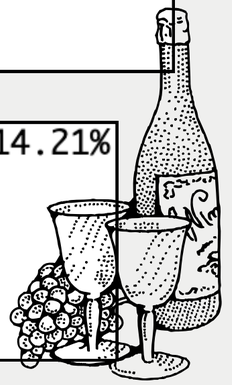
- **Class Error:** "Bad": Slightly higher error (13.18%). "Good": Reduced error (20.25%), improving minority class predictions.

- **Cluster Plot:** Shows clearer separation and better classification for "Good" wines.

- **Balanced Focus:** Model now performs better for "Good" wines, addressing the earlier bias toward "Bad" wines.




```
        OOB estimate of  error rate: 14.21%
Confusion matrix:
     Bad Good class.error
Bad  830  126   0.1317992
Good  33  130   0.2024540
```



Random Forest Model

Performance Metrics

	RESULT	EXPLANATION
Accuracy	82.9%	• Out of the total predictions, 82.9% matched the actual wine quality labels
Adjust Rand Index	0.32	• Moderate clustering agreement reflects the model's limitation for minority class "Good" predictions.
Confusion Matrix	Error rate: 16.67% (Bad)	• Among the "Bad" wines in the test data, 16.67% were incorrectly classified as "Good" wines.
	Error rate: 20.37% (Good)	• Among the "Good" wines in the test data, 20.37% were incorrectly classified as "Bad" wines.

The Random Forest model has proven effective at distinguishing between "Good" and "Bad" wines with an accuracy of 82.9%. This indicates that the model can reliably predict the quality of most wines in the dataset.

Random Forest Model

Summary Function

	Group.1	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
1	Bad	8.009836	0.5739754	0.2312568	2.518852	0.09119126
2	Good	9.085965	0.3892105	0.3965789	2.544737	0.07586842

	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates
1	16.92213	52.24454	0.9968152	3.328743	0.6346175
2	13.64035	35.09649	0.9961684	3.275263	0.7420175

	alcohol	quality	quality.label
1	10.15574	5.434426	1.030055
2	11.47515	6.315789	1.377193

Predicted Quality

	Group.1	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
1	Bad	8.261737	0.5419953	0.2649765	2.498944	0.08976761
2	Good	8.294444	0.4362037	0.3142593	2.730556	0.07007407

	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates
1	15.97770	48.73826	0.996787	3.312230	0.6499296
2	17.44444	43.70370	0.995672	3.346111	0.7405556

	alcohol	quality	quality.label
1	10.32113	5.464789	1
2	11.63642	7.055556	2

Actual Quality

3 Tops properties

- **Alcohol :**
  - Higher for “Good” wines than “Bad” in both
  - Good wines have significant higher alcohol levels, making it the most importance factor.
- **Sulphates:**
  - Higher for “Good” wines than “Bad” in both
  - Sulphates enhance flavor and preservation, strongly influencing wine quality.
- **Volatile Acidity:**
  - Lower for “Good” wines than “Bad” in both
  - Lower volatile acidity improve taste, making this a critical factor for high-quality wines.

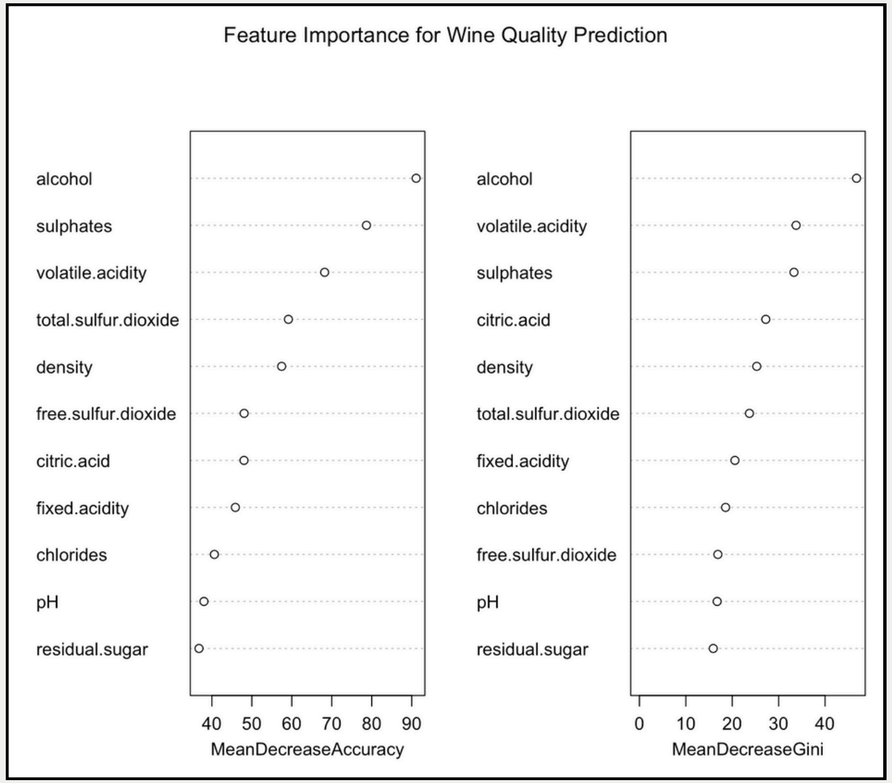
Random Forest Model

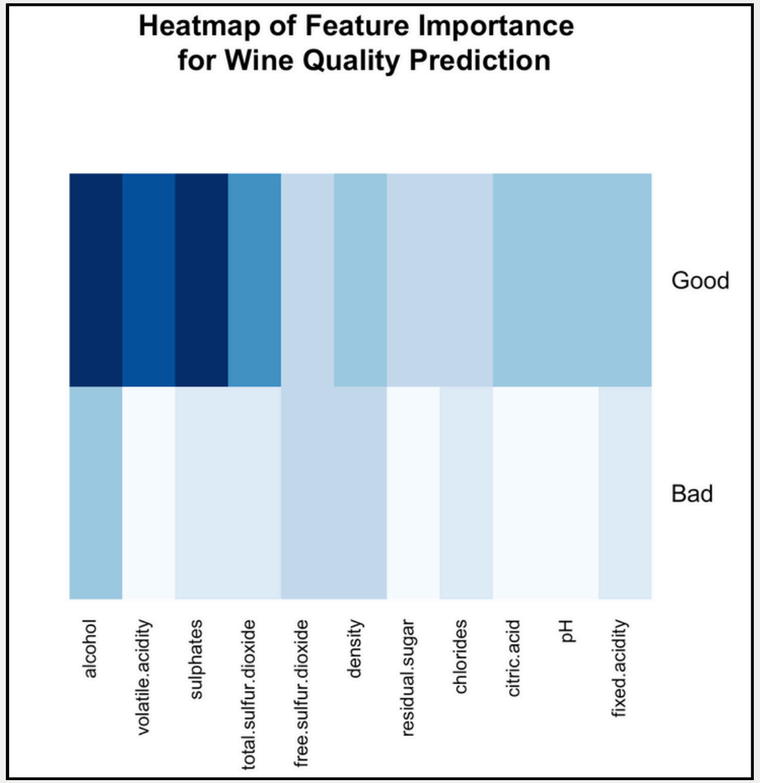
3 Tops properties

Variable Importance

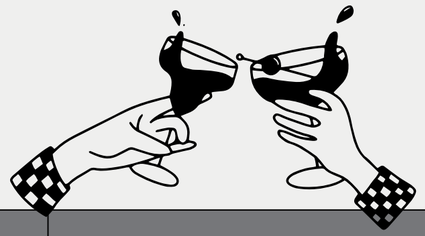
- The **heatmap** clearly shows that "Good" wines have higher alcohol and sulphates and lower volatile acidity compared to "Bad" wines.

- **The feature importance plots** (Mean Decrease Accuracy and Gini) confirm the rankings of alcohol, sulphates, and volatile acidity as the most influential factors.





Comparative Analysis



	NAIVE BAYES	RANDOM FOREST	EXPLANATION
Accuracy	80.83%	82.9%	Naive Bayes predicted 80.83% of wine labels, slightly lower than Random Forest.
Adjust Rand Index	0.21	0.32	Naive Bayes shows moderate agreement with actual labels, with room for improvement for "Good" class predictions.
Confusion Matrix	Error rate: 20.37% (Good)	Error rate: 20.37% (Good)	Both models show similar misclassification rates for "Good" wines, but Naive Bayes is slightly less accurate.
3 Top Features	Alcohol, Sulphates, Volatile Acidity	Alcohol, Sulphates, Volatile Acidity	Both models agree on Alcohol and Sulphates and Volatile Acidity as key features.

Key Takeaway

- **Random Forest outperforms Naive Bayes** in terms of accuracy and minority class performance.
- **Random Forest** also identifies important physicochemical properties, offering deeper insights for marketing and product decisions.

Business Implications and Recommendation

Implications

- **Model Effectiveness:**
  - **Naive Bayes Model:** With an accuracy of 80.83%, Naive Bayes performs well in classifying wine quality based on physicochemical features, offering a straightforward and efficient model for wine quality classification.
  - **Random Forest Model:** While slightly more accurate (82.9%), it requires more complexity. Naive Bayes still outperforms in terms of accuracy with minimal complexity and provides value in simpler scenarios.

- **Impact of Class Imbalance:**
  - Naive Bayes model is highly effective in handling imbalanced datasets, maintaining robustness even with skewed distributions between the “Bad” and “Good” wines.
  - The Random Forest model’s handling of class imbalance improves prediction accuracy for "Good" wines, but Naive Bayes provides a simpler and effective solution for classifying minority classes.

- **Feature Importance:**
  - Top Features Identified: **Alcohol, Sulphates, Volatile Acidity** were identified as the most important features in predicting wine quality.
  - These features highlight key areas winemakers can focus on to improve product quality, aligning production with quality characteristics demanded in the market.

Recommendations

For Wine Producers

- **Focus on Alcohol Content:** Since alcohol significantly influences wine quality, experimenting with different fermentation processes to adjust alcohol levels could help producers achieve the optimal balance for quality wine.
- **Increase Sulphates for Better Preservation:** Ensuring adequate sulphate content can enhance flavor, extend the shelf life, and align with consumer demand for higher-quality wines.
- **Reduce Volatile Acidity:** Maintaining low levels of volatile acidity will not only improve taste but also enhance the appeal of wines in the high-quality segment.

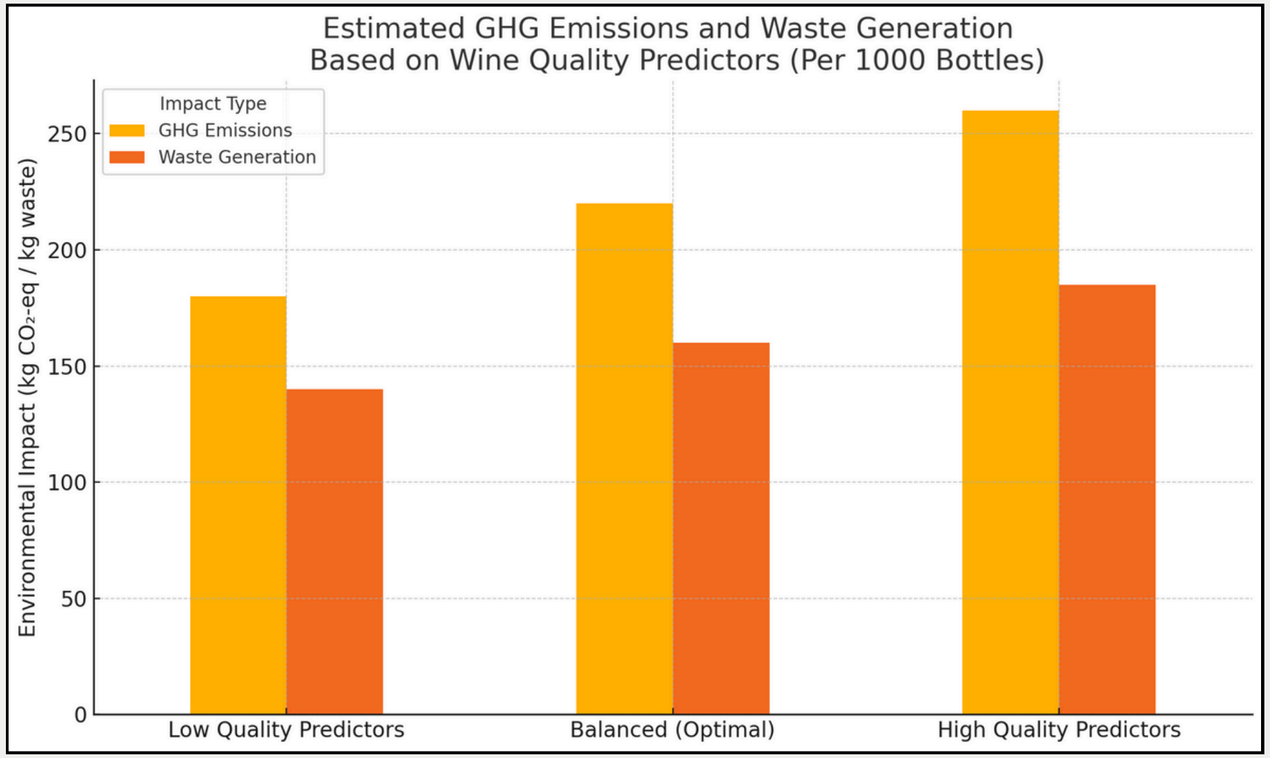
For Marketing and Product Decisions

- **Targeting "Good" Wines:** The insights from the models can guide marketing efforts by emphasizing the importance of alcohol, sulphates, and volatile acidity as key selling points for premium wine products.
- **Use Model Insights for R&D:** The Naive Bayes model can inform the development of wine by identifying key physicochemical traits that correlate with high-quality wines. This insight can be used to enhance product development and continuous improvement.

Data-Driven Marketing Campaigns

- Wine brands can utilize these insights in marketing campaigns by highlighting scientifically-backed quality factors, appealing to consumers who value quality consistency.

Optional: estimated GHG emissions and waste generation



A comparative bar chart showing estimated GHG emissions and waste generation based on the levels of key wine quality predictors—Alcohol, Sulphates, and Volatile Acidity:

Low Quality Predictors (lower alcohol, sulphates, higher acidity):
→ Lower environmental impact, but compromises wine quality.

Balanced (Optimal):
→ Offers a middle ground between sustainability and quality.

High Quality Predictors (higher alcohol, more sulphates, lower acidity):
→ Higher quality wines but also increased GHG emissions and waste due to more intensive processing, energy use, and chemical input.

Appendix

For more detailed information in part of coding, please access the full document by scanning the provided QR code with your mobile device's camera or a QR code scanning application.