White Wine Quality Prediction
Machine Learning Model

Veronika Angyalová
Digital Transformation Management
Alma Mater Studiorum
Università di Bologna
Email: veronika.angyalova@studio.unibo.it

Abstract—**This project investigates the potential of machine learning to automate quality prediction based on objective physicochemical properties. Using the White Wine Quality dataset from the UCI Machine Learning Repository, we analyzed wine samples characterized by 11 chemical features. Exploratory data analysis revealed a significant class imbalance, with the majority of wines receiving average scores. To address this, three distinct modeling approaches were implemented using Random Forest algorithms: (1) Binary Classification (good vs bad), (2) Multi-Class Classification (good/average/bad), and (3) Regression Analysis for exact score prediction. The Binary Radom Forest Classifier achieved the highest performance with an accuracy of 87.95% on a held-out test set. These results demonstrate that ensemble learning methods can effectively identify premium wine samples based solely on chemical composition, offering a scalable solution for industrial quality control.**

1. Introduction

Quality assessment plays a crucial role in wine market value and customer satisfaction. Traditionally the wine quality is evaluated by experts through sensory analysis. This project is aiming to create a prediction model that would help with predicting the wine quality based on its chemical characteristics.

The study utilized the White Wine Quality dataset from the UCI Machine Learning Repository, which contains 4,898 wine samples from the north of Portugal. Each sample is characterized by 11 chemical features, including acidity levels, residual sugar, alcohol content and quality scores. The core objective of this study is to develop a predictive framework that evaluates wine quality through three increasing levels of detail:

1) **Binary Classification** – distinguishing between premium and non-premium samples (“Good” vs. “Bad”).
2) **Multi-Class Classification** – introduction of an “Average“ category to separate mid-range wines from the extremes.
3) **Regression Analysis** – predicting the exact quality scores based on the wine chemical profile.

The integration of machine learning into the wine industry represents a significant shift toward data-driven quality control. By automating the assessment of chemical profiles, producers can identify potential quality issues early in the production cycle and reduce the reliance on costly and time-consuming professional tastings.

1.1. Project Scope and Methodology

The author managed the entirety of the project’s technical architecture, from initial data ingestion to final model validation. To ensure the integrity of the findings, the development process followed a structured four-phase methodology:

1) **Exploratory Data Analysis:** Conducted initial data profiling to identify underlying label distributions. This included generating correlation matrixes and visualizations to isolate the most significant chemical drivers of wine quality, such as alcohol content and density.
2) **Data Preprocessing and Engineering:** Addressed the inherent class imbalance of the dataset by defining specific target boundaries. The data was meticulously partitioned into training and testing sets, and the target variable was transformed to support three distinct algorithmic approaches.
3) **Model Development and Training:** Configured and trained Random Forest models for classification and regression tasks. This phase required utilizing ensemble methods to ensure the models learned underlying chemical patterns rather than simply memorizing the training data.
4) **System Validation:** Designed and implemented a manual predictive testing framework. This system allowed for the input of raw chemical data to simulate real-world vineyard testing, verifying the practical accuracy and robustness of the trained models against the ground truth labels.

This systematic approach ensures that the findings presented in this report are both reproducible and directly applicable to quality control systems.

2. Proposed Method

2.1. Dataset Acquisition and Initial Exploration

The foundation of this project is the White Wine Quality dataset form the UCI Machine Learning Repository. The data was imported via the Pandas library, then the initial exploration was conducted using `.info()` and `.describe()` methods. This helped us better understand the contents of the dataset. The dataset didn't contain any null values, which was checked using `.isnull().sum()`. We discovered that there was no need for additional data cleaning. To better understand the target variable, a distribution analysis was performed. As shown on the **Figure 1**, the quality scores are not evenly distributed. The vast majority of samples are concentrated in the mid-range, scores 5 and 6, while extreme scores like 3 and 9 are significantly underrepresented. This observation suggests that a model might struggle to accurately predict rare quality scores due to a lack of training examples at the edges of the scale.

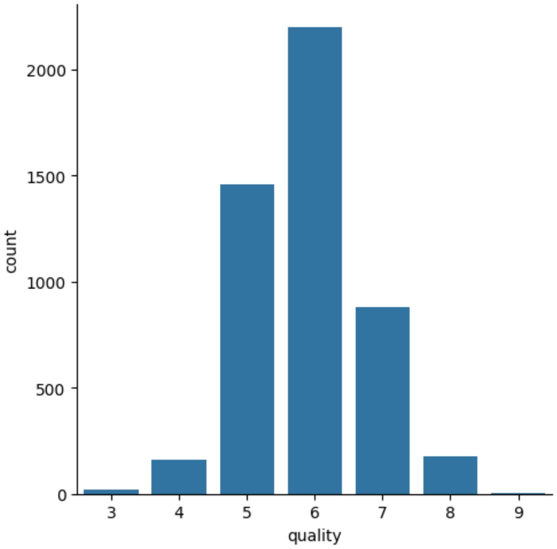

Figure 1. Distribution of wine quality scores (3–9).

2.2. Feature Engineering and Correlation Analysis

To identify which chemical properties affect quality, corelation analysis was conducted. Firstly, bar plots were used to visualize each individual psychochemical variable against the quality score. These plots represent the mean value of a chemical for each quality score (3 – 9). For instance, chlorides (**Figure 2**) showed a consistent downward trend as its concentration decreased, the quality of the wine generally increased. Similar visual trends were noted for Citric Acid (positive trend), providing a visual foundation for the predictive power of the dataset.

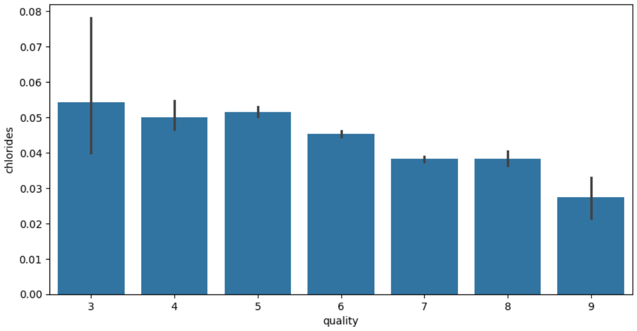

Figure 2. Mean chloride concentration across different quality levels.

To quantify these relationships more precisely, corelation matrix was generated using the Seaborn heatmap (**Figure 3**). This matrix calculates the linear relationship between every pair of variables on a scale from -1 to $+1$. By examining the "Quality" column in the heatmap, the key drivers of the model were identified:

Positive correlations:

- *Alcohol (0.44):* Emerged as the most significant positive driver. This indicated that wines with higher alcohol content are associated with higher quality ratings in this dataset.

Negative correlations:

- *Density* (-0.31)*:* This measure had the strongest negative correlation in the dataset. Density is heavily influenced by the amount of dissolved sugar and alcohol. In this dataset, higher density generally indicates higher residual sugar, which experts often associated with lower quality white wines.
- *Chlorides* (-0.21)*:* This essentially measures the amount of salt in the wine. A higher correlation here suggests that salty undertones are undesirable and lead to lower scores.
- *Volatile acidity* (-0.19)*:* This measures the amount of acetic acid. At higher levels, it gives wine an unpleasant, vinegar-like taste and smell. The negative correlation confirms that its presence lowers its quality ratings.

By identifying these connections, the project moved from just looking at raw data to the modeling phase. This correlation analysis helped validate the integrity of the data before applying the Random Forest algorithm. Because Random Forest models inherently prioritize the most influential variables, these identified correlations—such as alcohol and density—confirmed that the model was focusing on the most important data points. This ensures that the final predictions are grounded in the actual chemical logic of the samples, making the results more accurate and easier to interpret.

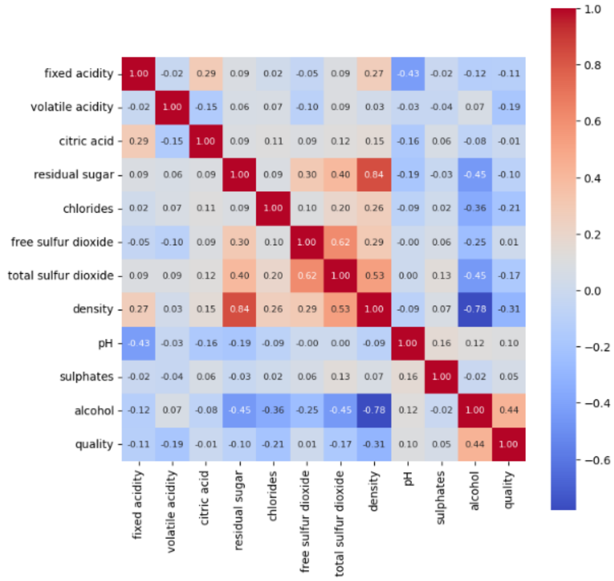

Figure 3. Correlation matrix showing the relationships between physico-chemical properties.

2.3. Data Preprocessing and Partitioning

To address the distribution challenges identified in Section 2.1, the study was organised into three distinct experimental frameworks:

- **1. Binary Classification:** The target was simplified into a "Pass/Fail" system ("Good" ≥ 7, "Bad" < 7). This maximizes reliability by focusing on identifying premium wines.
- **2. Multi-Class Classification:** A 3-tier approach was established to separate the mid-range "Average" wines (score 6) from the high and low extremes.
- **3. Continuous Regression:** The model was tasked with predicting the exact numerical score. While the training data exists between 3 and 9, the regression approach allows for a continuous output across the full quality scale.

Once the exploratory analysis was complete, the dataset was prepared for the machine learning model. This process involved separating the data into 2 parts:

- **X (features):** The original wine dataset was modified by dropping the quality column. This resulted in a matrix of 11 chemical variables.
- **Y (target):** The quality column was isolated. Depending on the experiment, this column was transformed using specific `lambda` functions (Binary, Multi-Class) or kept in its raw form (Regression).

To properly grade the model, the data was partitioned using the `train test split` function. Using a standard 20% split, the original dataset of 4,898 wines was divided into 2 groups:

- **Training set (3,918 wines):** This is the data the model learns on to find patterns between chemistry and quality.
- **Testing set (980 wines):** These are wines the model has never seen before. We used this group to test the model and see if it actually learned, or if it was just memorising the answers.

To ensure that results remain consistent across multiple executions, a `random state=2` parameter was applied. This acts as a fixed seed for the random split.

2.4. Model Training

For the classification task, the **Random Forest Classifier** was selected. This model was chosen because it is an ensemble method, meaning it builds multiple decision trees and merges their results together. This approach is superior to a single decision tree because it reduces the risk of the model simply memorizing the training data and instead helps it learn general patterns. For the regression task, the **Random Forest Regressor** was utilized, which operates on the same principle but predicts continuous numerical value by averaging the outputs of the trees rather than voting on a category. The training was performed using the `.fit()` function. During this phase, the model analysed the wines in the training set, comparing their 11 chemical properties against the target labels. By the end of this process, the model had developed a set of internal rules to determine wine quality based on the chemical signatures discovered during the training.

3. Results

3.1. Experimental Results and Evaluation

After training the Radom Forest models on the training samples, they were asked to predict quality of the 980 wines from the test set.

3.1.1. Binary Classification Results. The binary model achieved the highest performance with an Accuracy Score of 87.95%. This means that for every 100 wines the model analysed, it correctly identified whether they were "Good" or "Bad" approximately 88 times. In the context of wine quality, which is usually a subjective human taste, this is considered a very high performance.

3.1.2. Multi-Class Classification Results. When the task was made more granular (adding "Average" category for score 6), the Accuracy Score achieved was 69.08%. This decrese is expected because the chemical difference between a "Bad" wine (score 5) and "Average" wine (score 6) is very subtle, making the classification boundary harder to define than the simple Good/Bad spit.

3.1.3. Regression Results. The regression model achieved a Mean Squared Error (MSE) of 0.40 and R2 Score of 0.48. An MSE of 0.40 implies that on average the model's prediction deviated form the actual human score by approximately 0.63 points. This confirms that while exact prediction is difficult due to subjectivity, the model is consistently close to the expert ratings. To verify the models in real-world scenarios, I created a predictive system where I manually inputted the chemical data of specific wines to see if the output matched the original dataset.

3.2. Deeper Insights into Model Behavior

The Impact of Class Imbalance: The drop in accuracy from 87.95% (Binary) to 69.08% (Multi-Class) underscores the complexity of sensory modeling, specifically the lack of distinct thresholds defining the ”Average” tier. Because the dataset is heavily imbalanced toward scores of 5 and 6, the Random Forest model struggled to distinguish the subtle chemical thresholds that elevate a wine from 5 (Bad) to 6 (Average). The model is highly confident when separating extreme scores, but the chemical overlap in the middle tiers causes expected confusion.

Feature Importance: One of the key advantages of using a Random Forest algorithm is its ability to rank feature importance. Consistent with our initial exploratory data analysis, the model relied heavily on Alcohol and Density to make its decision splits. Volatile acidity also played a crucial role in penalizing lower-quality wines. This confirms that the machine learning model didn't just find random mathematical noise but it successfully learned and applied the exact same chemical rules that human sommeliers would use to judge wine structure.

3.3. Validation using Predictive System

To verify the practical applicability of the study a manual predictive system was developed to test distinct chemical inputs against all three trained models.

3.3.1. Classification Validation (Binary and Mulit-Class). Both classifications models were tested using inputs corresponding to known quality tiers. The Binary model successfully distinguished between high-quality and low-quality samples, correctly assigning class 1 (“Good”) and class 0 (“Bad”) respectively. The Multi-Class model was subjected to a broader test, correctly mapping inputs to the “Bad”, “Average” and “Good” categories. This confirmed that the decision boundaries between the three tiers were functioning as intended.

3.3.2. Regression Validation. The Regression model operates differently, outputting a continuous floating-point value (e.g. 5.87) rather than a fixed category. Since the original dataset uses standard integer scores (3, 4, 5, etc.) a post processing step was added to round the model's output to the nearest whole number. In our test, the model consistently produced floating point predictions close enough to the actual targets that when rounded they matched the true integer scores (prediction of 5.87 was correctly rounded to 6).

4. Conclusion

This project successfully demonstrated that Machine Learning can be used to objectively access white wine quality. The Random Forest Classifier (Binary) proved to be the most reliable tool for quality control, achieving an accuracy of 87.75%.

While the Regression and Multi-Class models provided more granular detail, they highlighted the inherent difficulty in chemically distinguishing between “Average” and slightly above average wines. The regression analysis showed that the model predictions are typically within 0.63 points of the human score, which is a strong baseline. Ultimately, this study confirms that physicochemical analysis, particularly of alcohol and volatile acidity, offers a robust and scalable alternative to traditional sensory testing.

4.1. Limitations and Future Research

Despite the strong predictive results, this study has certain limitations. The dataset is restricted to white ‘’Vinho Verde” wines from Portugal. Therefore, the specific chemical thresholds learned by this model may not perfectly generalize to red wines or wines grown in different climates, such as California or France. Additionally, the severe class imbalance limited the model's ability to confidently predict rare, ultra-premium wines (scores of 9 or 10). Future work could potentially focus on addressing this imbalance.