“Predicting Wine Quality Using Machine Learning Techniques: A Data Science Lifecycle Approach”

line 1: 1st Given Name Surname   
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 4th Given Name Surname  
line 2: *dept. name of organization*  
*(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 2nd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 5th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCIDline 1: 3rd Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

line 1: 6th Given Name Surname  
line 2: *dept. name of organization   
(of Affiliation)*  
line 3: *name of organization   
(of Affiliation)*line 4: City, Country  
line 5: email address or ORCID

***Abstract*—This paper presents a data science project aimed at predicting wine quality based on physicochemical attributes. Using the Wine Quality dataset from the UCI Machine Learning Repository, we applied a full data science lifecycle—from data acquisition and preprocessing to exploratory analysis, feature engineering, and machine learning modelling. Our approach employs the CRISP-DM framework. A Random Forest regression model was implemented and evaluated using RMSE, yielding promising results. Limitations and potential improvements are discussed alongside considerations of ethical, legal, and sustainability aspects.**

**Keywords— Wine Quality Prediction, Machine Learning Modeling, Feature Engineering, Data Cleaning, Statistical analysis, Data Visualization.**

1. **Introduction:** 
   1. **The Role of Machine Learning in Wine Quality Assessment**

The assessment of quality plays an increasingly vital role in the food and beverage industry, impacting consumer satisfaction, brand loyalty, and adherence to regulatory standards. Maintaining consistent quality is not merely a matter of preference but a fundamental requirement for businesses to thrive in a competitive market and to meet the expectations of discerning consumers. Furthermore, the industry is subject to a complex web of regulations aimed at ensuring product safety and quality, making reliable assessment methods paramount.

Machine learning offers a transformative potential to augment and enhance quality assessment in the food and beverage industry by providing objective, efficient, and scalable solutions. By analyzing large datasets of quantifiable parameters, such as physicochemical properties, machine learning algorithms can learn to predict sensory quality attributes with a high degree of accuracy. The application of artificial intelligence (AI), which encompasses machine learning, can lead to automated systems capable of processing vast amounts of data to optimize production processes and maintain consistent product quality across different batches.

**1.2 Aim and Objectives**

The primary aim of this project is to develop a reliable machine learning model for predicting red wine quality from physicochemical features. To achieve this aim, we set out the following objectives:

* **Data Acquisition:** Obtain a suitable open dataset of wine samples with quality labels.
* **Data Preprocessing:** Clean and prepare the data (handle missing values, outliers, etc.) to ensure it is analysis ready.
* **Exploratory Data Analysis (EDA):** Investigate the dataset to discover trends, correlations, and distributions among features and the target quality score.
* **Feature Engineering:** Create or transform features (if needed) to improve the model’s predictive performance (e.g., by capturing non-linear relationships or interactions).
* **Model Development:** Implement a regression model (chosen based on literature and dataset characteristics) and tune it for optimal performance.
* **Evaluation:** Evaluate the model using appropriate metrics (such as RMSE) and validate its performance on unseen data.
* **Ethical Considerations:** Ensure the project adheres to legal, ethical, and professional standards, including responsible use of data and transparency in model interpretation.

By fulfilling these objectives, we seek to demonstrate an end-to-end solution for wine quality prediction and to gain insights into which chemical factors most strongly influence quality ratings**.**

1. **Methods**

**A. Data Collection and Dataset Description**

We utilized the Wine Quality dataset (red wine variant) from the UCI Machine Learning Repository. This open dataset contains 1,599 samples of Portuguese “Vinho Verde” red wine, each with 11 measured physicochemical attributes and a quality score (sensory rating) provided by wine experts on a scale from 0 (very poor) to 10 (excellent). In practice, the quality ratings in this dataset range from 3 to 8, with most wines rated in the middle of this range. The physicochemical features include metrics such as acidity levels, sugar content, pH, alcohol percentage, and concentrations of sulphur dioxide.

Given that the dataset is public and well-documented, we confirmed that there were no missing values present. All 1,599 samples had complete entries for each of the 11 features. This eliminated the need for imputation of missing data. We did, however, examine the data for any inconsistent entries or obvious data errors.

**B. Data Preprocessing and Exploratory Data Analysis**

In the Data Preprocessing phase, we performed basic cleaning and formatting of the dataset. Since the data came ready in a CSV format, the main preprocessing steps included: ensuring correct data types for each feature (all features are numeric, with quality as an integer), normalizing naming conventions, and checking for duplicate records (none were found). Because our chosen modelling algorithm (Random Forest) is not sensitive to feature scaling, we did not need to standardize or normalize the numeric features for the model’s sake; however, for certain analyses like correlation, all features being numeric allowed direct computation without additional transformation.

**For Exploratory Data Analysis (EDA),** we used statistical summaries and visualizations to understand the data’s characteristics. We calculated descriptive statistics for each feature (mean, median, standard deviation, min, max) to get a sense of typical values and variability. The quality scores had a mean around 5.6, indicating that most wines are of average quality, and a relatively symmetric distribution around 5–6. We plotted the distribution of the quality variable, which showed that ratings 5 and 6 were the most frequent, with fewer wines rated very low (3 or 4) or very high (7 or 8).

**3. Feature Engineering**

Based on the findings from EDA, we performed feature engineering to potentially enhance the model’s predictive power. The goal was to create new features or transform existing ones to capture aspects of the data not immediately apparent to the base model. We employed several strategies:

* **Interaction Features**: We introduced an interaction term between alcohol and volatile acidity (Alcohol × Volatile Acidity) to account for the combined effect of these two influential factors. The intuition is that the impact of acidity on quality might vary at different levels of alcohol; for instance, a wine with high acidity might be better received if it also has higher alcohol to balance the flavour. Similarly, we experimented with an interaction between fixed acidity and pH (Fixed Acidity × pH), since pH is essentially a measure of overall acidity.
* **Polynomial Features**: To capture potential non-linear relationships, we added a squared term for alcohol content (Alcohol²). This can help model any diminishing or increasing returns effect of alcohol on quality (e.g., perhaps extremely high alcohol content might not continue to improve quality linearly).
* **Binning**: We created binned versions of certain continuous features to reduce noise and possibly capture threshold effects. For example, **residual sugar** was binned into categories "dry", "off-dry", and "sweet" based on typical wine sweetness thresholds (dry wines having <4 g/L residual sugar, etc.). These categories might align better with how human tasters perceive sweetness levels, which could improve the model if the relationship between sugar and quality is non-linear
* **Ratio Feature**: We added a new feature as the ratio of free sulphur dioxide to total sulphur dioxide (FreeSO2/TotalSO2). This ratio is related to the proportion of sulphites available to act as preservatives (free form) versus bound forms.

After creating these new features, we evaluated their usefulness. We checked correlation of the engineered features with quality and found, for instance, that the Alcohol × Volatile Acidity interaction had a slight positive correlation with quality.

**4. Model Development and Evaluation**

For the predictive modelling, we selected a **Random Forest regression** algorithm as our primary model. The choice was motivated by several considerations. First, as an ensemble of decision trees, Random Forests are known for their ability to capture non-linear relationships and interactions between features automatically. This suits our problem where relationships (like those between chemical properties and quality) can be complex. Second, Random Forests offer robust performance even without extensive parameter tuning and are resistant to overfitting due to the averaging of many decision trees and the use of bootstrap sampling and feature randomness.

We implemented the Random Forest using the scikit-learn library in Python (Random Forest Regressor). The training set (approximately 1,279 samples after the 80/20 split) was used to train the model. Initially, we trained a baseline Random Forest with default parameters (100 trees, no maximum tree depth constraint, using the square-root of feature count for features considered at each split, etc.). To ensure the model’s performance was not dependent on a lucky split, we employed a 5-fold cross-validation on the training data when comparing different parameter settings. We then performed hyperparameter tuning to further improve performance: using grid search, we tried varying the number of trees in the forest (n\_estimators tested values 100, 200, 500) and the maximum depth of the trees (e.g., None vs. limiting to 5, 10, or 15 levels).

For evaluation, the primary metric used was **Root Mean Squared Error (RMSE)**, which measures the average prediction error in the same units as the target (quality points). RMSE is defined as the square root of the mean of squared differences between predicted quality values and actual quality values. We chose RMSE because it penalizes larger errors more severely than smaller ones, which is useful in our context since a few large mispredictions (for example, a wine truly rated 8 predicted as a 5) would be particularly undesirable. Additionally, we reported the **R-squared (R²)** value of the model on the test set, which represents the proportion of variance in wine quality that is explained by the model.

1. **Results**

After training the Random Forest regression model on the wine dataset and performing the above procedures, we obtained the following results:

* **Descriptive Statistics of Quality**: The quality scores in the test set had a similar distribution to the training set (most wines rated 5 or 6, with few extremes). The baseline expectation (predicting the mean quality for all instances) would yield a certain error as a reference. The standard deviation of quality in the whole dataset is about 0.88, which means a naive predictor would have an RMSE around 0.88 if it always predicted the mean (~5.6). Our model’s performance can be compared against this baseline.
* **Exploratory Analysis Findings**: The EDA confirmed that a few features play a significant role in wine quality. For example, wines in the test set with alcohol content above 12% were predominantly predicted (and observed) to have quality ratings of 6 or higher, whereas those with very high volatile acidity (>1.0 g/L acetic acid) were often predicted to be of lower quality.
* **Model Performance**: The Random Forest regression model achieved an **RMSE of approximately 0.57** on the test set (in quality score units). This means on average the prediction error was about half a point on the wine quality scale, which is a substantial improvement over the baseline error (~0.88) one would get by predicting the average quality for all wines. In practical terms, an error of 0.57 indicates that the model’s predictions are typically within about ±0.57 of the true quality rating. Considering quality is an ordinal score and humans themselves might have some variability in scoring, this level of accuracy is quite promising.
* **Feature Importance**: Analysing the trained Random Forest model, we extracted the feature importance values. These reflect how much each feature (including any engineered features) contributed to reducing prediction error across the trees. The results reinforced our expectations from the correlation analysis. **Alcohol content** was the top predictor of quality by a considerable margin. **Volatile acidity** and **sulphates** were also among the most important features, roughly tying for the second place in importance. The interaction feature Alcohol × Volatile Acidity received a modest importance, suggesting that while the model found it somewhat useful, the primary individual effects of alcohol and acidity were more significant. Features like **density** and **chlorides** had lower importance, implying they added little incremental predictive power when the other attributes were already considered. Interestingly, the **free sulphur dioxide to total sulphur dioxide ratio** feature we engineered had a small but non-negligible importance, which could indicate that wines with a higher free-to-total SO₂ ratio (i.e., more effective preservatives) tended to be predicted slightly higher in quality, after accounting for other factors.
* **Comparison of Engineered Features**: To assess the impact of our feature engineering, we also trained a version of the Random Forest model without the newly engineered features. The model with engineered features delivered a slightly better performance (a few hundredths lower RMSE) than the one without, indicating a minor benefit. For example, including the Alcohol × Volatile Acidity interaction and the sulphates feature together allowed the model to predict quality for some edge-case wines more accurately than relying on each feature separately.

Overall, the results demonstrate that our approach can effectively predict wine quality to a useful degree. The model’s performance is **promising** for a regression task on subjective quality ratings, and the identified important features align well with domain knowledge of oenology. In the next section, we discuss the implications of these findings, the limitations of our current model, and possible avenues for further improvement.

1. **Discussion**

**Limitations and Future Work**

While the Random Forest model achieved reasonable accuracy in predicting wine quality, there are several limitations to note. **Firstly**, the model is inherently limited by the information in the dataset. The physicochemical features explain roughly half of the variance in quality; the remaining variance might be due to factors not captured in the data (such as grape variety, fermentation process details, aging, or even subjective differences in human tasters’ preferences). **Secondly**, the dataset itself is of moderate size (1599 samples for red wine). A larger dataset covering more types of wine and a broader range of quality scores could help train more generalizable models. The relatively small sample size also restricted the complexity of models we could reliably train without overfitting.

Another limitation relates to how **quality is measured**. The quality score is an average of at least three judges’ ratings in the original data. There is an inherent subjectivity and possible inconsistency in these ratings. Our model treats quality as a precise numerical target, but two wines with the same chemical profile might be scored differently on a different day or by different experts

In terms of the **modelling approach**, while Random Forest was a strong choice, there is room to try alternative or more advanced techniques. **Future work** could include exploring ensemble methods like Gradient Boosting Machines (e.g., XGBoost or LightGBM) which often achieve state-of-the-art results in structured data problems and might capture subtle patterns with more nuanced weighting of trees.

Another avenue for improvement is **hyperparameter tuning and model optimization**. Our grid search was relatively coarse due to time and computational constraints. A more extensive search (or using advanced techniques like Bayesian optimization) might find a better combination of parameters (e.g., an optimal number of trees beyond 200, or different max features setting).

Finally, incorporating additional data attributes could improve predictions. For example, metadata such as the **wine variety or vineyard** could account for quality factors not captured by chemistry alone. If such data were available, merging it with the current feature set could increase the explanatory power of the models. We might also look at the **white wine dataset** (which is separate but similar) and see if a combined model or comparative analysis yields new insights (though quality ratings between red and white might not be directly comparable, a multi-task model could be an interesting exploration).

1. **Ethical and Societal Considerations**

Throughout the project, we remained mindful of **ethical, legal, and professional considerations**. One aspect is data usage and privacy: the dataset used is openly available and **contains no personal or sensitive information**, so privacy concerns are minimal. We ensured compliance with the dataset’s license and cited the source [2] appropriately. In a professional context, all team members’ contributions were documented, and we adhered to academic integrity guidelines by avoiding plagiarism and acknowledging sources and prior work.

**Transparency** in the model’s decision-making is important for trust. While ensemble models like Random Forest are not the most interpretable, we mitigated this by analysing feature importances and explaining which factors drive predictions. There is also a risk of **over-reliance or misuse** of such a model. If a winery were to use a predictive model to grade wines, it should not fully replace human sommeliers or quality experts without thorough validation, as it could undermine the nuanced expertise those professionals provide.

In terms of **social impact**, the adoption of ML for wine quality could affect jobs (sommeliers, quality inspectors) and market dynamics. We should consider the balance between technological assistance and human expertise. Ideally, such models can help wine makers improve their process (e.g., by identifying key factors to monitor) rather than replace human judgement. Ensuring **fairness** is also part of ethical practice: although not directly applicable to wine (since we are not dealing with protected human attributes), fairness here could relate to not favouring certain wine varieties or producers unjustly.

From a **sustainability** standpoint, our project had a modest computational footprint (Random Forest training on 1.6k samples is not resource-intensive). However, scaling this up or using heavier algorithms means considering energy efficiency

In summary, we conducted this project with a commitment to responsible data science. We have demonstrated the technical feasibility of predicting wine quality, but any real-world implementation would need to address transparency, avoid misuse, and complement (rather than supplant) human judgement in the loop.

1. **Conclusion**

In this study, we demonstrated an end-to-end data science approach for predicting wine quality using machine learning techniques, following an IEEE-compliant structure and the CRISP-DM lifecycle. We successfully developed a Random Forest regression model that can predict the quality of red wine with reasonable accuracy using only physicochemical features. The comprehensive exploration of the data revealed which chemical properties most strongly influence quality, aligning with domain knowledge (for example, alcohol content and acidity are critical factors). By engineering additional features and carefully tuning the model, we achieved an RMSE of around 0.6 on the test set, indicating that the model’s predictions are typically within half a point of the true expert rating.

The approach highlights the value of combining domain expertise (oenology and chemistry) with data science techniques. Each phase of the project—from data cleaning and EDA to modelling and evaluation—contributed to the outcome, illustrating the importance of the data science lifecycle in developing robust solutions.

In conclusion, the project demonstrates that machine learning, guided by a structured process and combined with expert knowledge, can effectively model wine quality. The results offer useful insights for wine producers into which factors contribute most to quality and lay the groundwork for intelligent decision-support systems in quality control. Future work and deployment should proceed with a balance of optimism for technological benefits and caution for maintaining fairness, sustainability, and trust in the outcomes.

**9. References**

1. How is AI Enhancing Quality Assurance in the Food and Beverages Industry? - KnowHow, <https://knowhow.distrelec.com/food-and-beverages/how-is-ai-enhancing-quality-assurance-in-the-food-and-beverages-industry/>

2. The Impact of AI in the Food and Beverage Industry: An Expert Q&A With Benny Dor and Chad Merck - Aptean.com, <https://www.aptean.com/en-US/insights/blog/expert-qa-ai-in-food-and-beverage>

3. (PDF) Leveraging artificial intelligence and advanced food processing techniques for enhanced food safety, quality, and security: a comprehensive review - ResearchGate, <https://www.researchgate.net/publication/387939314_Leveraging_artificial_intelligence_and_advanced_food_processing_techniques_for_enhanced_food_safety_quality_and_security_a_comprehensive_review>

4. Modeling wine preferences by data mining from physicochemical properties - ResearchGate, <https://www.researchgate.net/publication/222430341_Modeling_wine_preferences_by_data_mining_from_physicochemical_properties>

5. Wine Quality Data Set, <https://www.stat.cmu.edu/~brian/valerie/617-2022/617-2021/project01/UCI%20ML%20data%20sets/Vinho%20Verde/UCI%20Machine%20Learning%20Repository_%20Wine%20Quality%20Data%20Set.pdf>

6. AI in Food Manufacturing: How Artificial Intelligence is Transforming Food Production, <https://integrio.net/blog/ai-in-food-manufacturing>

7. Review of AI-Powered Food Processing: Enhancing Safety and Sustainability, <https://www.researchgate.net/publication/385039096_Review_of_AI-Powered_Food_Processing_Enhancing_Safety_and_Sustainability>

8. What Are The 5 Steps in Data Science Lifecycle | Institute of Data, <https://www.institutedata.com/us/blog/5-steps-in-data-science-lifecycle/>

9. Data Science Life Cycle (All Steps Explained) in 2025 - Fynd Academy, <https://www.fynd.academy/blog/data-science-life-cycle>

10. Data Science Life Cycle: Detailed Explanation - OdinSchool, <https://www.odinschool.com/blog/data-science-life-cycle-detailed-explanation-2023>

11. 5 Key Steps in the Data Science Lifecycle Explained, <https://www.worlddatascience.org/blogs/5-key-steps-in-the-data-science-lifecycle-explained>