**Red Wine Quality Prediction Project**

**Project Description**

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).  
  
This dataset can be viewed as classification task. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

**Attribute Information**

Input variables (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 - alcohol  
Output variable (based on sensory data):  
12 - quality (score between 0 and 10)

What might be an interesting thing to do, is to set an arbitrary cutoff for your dependent variable (wine quality) at e.g. 7 or higher getting classified as 'good/1' and the remainder as 'not good/0'.  
This allows you to practice with hyper parameter tuning on e.g. decision tree algorithms looking at the ROC curve and the AUC value.

You need to build a classification model.

Inspiration

Use machine learning to determine which physiochemical properties make a wine 'good'!

**Dataset Link-**

<https://github.com/FlipRoboTechnologies/ML-Datasets/blob/main/Red%20Wine/winequality-red.csv>

# Predicting Red Wine Quality: A Machine Learning Approach

## 1. Problem Definition

### Background

Wine quality is a major determinant in the wine industry, influencing both consumer satisfaction and market value. For producers, maintaining high quality and consistency is crucial for brand reputation and profitability. The ability to predict the quality of wine based on its physicochemical properties can offer significant advantages. By understanding which factors most strongly impact quality, winemakers can refine their production processes, enhance quality control, and tailor their products to consumer preferences.

### Objective

The goal of this project is to develop a machine learning model that predicts the quality of red wine based on its physicochemical attributes. Utilizing historical data from wine samples, the objective is to create a predictive model that can accurately forecast wine quality ratings. This predictive capability could assist wineries in optimizing their production processes, improving quality control, and ensuring a consistent product offering.

## 2. Data Analysis

### Dataset Source

The dataset used in this project is sourced from a collection of red wine samples, encompassing a variety of physicochemical attributes and their corresponding quality ratings. This dataset is essential for training and evaluating our machine learning models. It provides a comprehensive view of the factors that contribute to wine quality and serves as the foundation for our analysis and model development.

### Structure

The dataset includes the following features:

* **Fixed Acidity:** Concentration of non-volatile acids.
* **Volatile Acidity:** Concentration of volatile acids, which can affect the taste of the wine.
* **Citric Acid:** Impacts the freshness and flavor of the wine.
* **Residual Sugar:** Indicates sweetness levels.
* **Chlorides:** Concentration of salt, affecting the taste and preservation.
* **Free Sulfur Dioxide:** A measure of the free form of sulfur dioxide, which helps in wine preservation.
* **Total Sulfur Dioxide:** Total amount of sulfur dioxide present.
* **Density:** Mass per unit volume of the wine.
* **pH:** Acidity level of the wine.
* **Sulphates:** Compounds that can influence taste and stability.
* **Alcohol:** Alcohol content by volume.
* **Quality:** Target variable representing the quality rating of the wine (scaled from 0 to 10).

### Key Characteristics

The dataset contains a diverse range of samples, allowing for a robust analysis of how different physicochemical properties impact wine quality. Key characteristics include:

* **Distribution of Quality Ratings:** The quality ratings are distributed between 3 and 8, with a majority of samples falling within the 5-7 range. This distribution provides a balanced view of the various quality levels and helps in understanding the factors influencing higher or lower ratings.
* **Feature Correlations:** Initial analysis reveals that certain features, such as alcohol content and volatile acidity, show strong correlations with wine quality. These correlations suggest that these features are significant predictors of wine quality.

## 3. Data and Its Interpretation

### Dataset Overview

The dataset consists of 1,599 red wine samples with 12 features. The target variable is the quality rating, which is an integer between 3 and 8. The features are numerical, and the dataset is structured as follows:

| **Fixed Acidity** | **Volatile Acidity** | **Citric Acid** | **Residual Sugar** | **Chlorides** | **Free Sulfur Dioxide** | **Total Sulfur Dioxide** | **Density** | **pH** | **Sulphates** | **Alcohol** | **Quality** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 7.4 | 0.70 | 0.00 | 1.9 | 0.076 | 11 | 34 | 0.9978 | 3.51 | 0.56 | 9.4 | 5 |
| 7.8 | 0.88 | 0.00 | 2.6 | 0.098 | 25 | 67 | 0.9968 | 3.20 | 0.68 | 9.8 | 5 |
| 7.8 | 0.76 | 0.04 | 2.3 | 0.092 | 15 | 54 | 0.9970 | 3.26 | 0.65 | 10.0 | 5 |

### Feature Analysis

#### Fixed Acidity

* **Interpretation:** Fixed acidity is a measure of non-volatile acids, which contribute to the wine’s overall acidity and taste. Higher fixed acidity can lead to a more acidic taste, potentially impacting the quality perception.
* **Distribution:** The fixed acidity feature varies from 4.6 to 15.9 g/dm³, with a mean value around 7.4 g/dm³. Wines with higher fixed acidity tend to have a more pronounced acidic taste.

#### Volatile Acidity

* **Interpretation:** Volatile acidity represents the concentration of volatile acids, which can impart a vinegary taste if excessively high. Lower volatile acidity is generally associated with higher quality wines.
* **Distribution:** This feature ranges from 0.12 to 1.58 g/dm³, with a mean value of 0.65 g/dm³. High volatile acidity is often correlated with lower quality ratings.

#### Citric Acid

* **Interpretation:** Citric acid enhances the freshness and flavor profile of the wine. It contributes to the overall balance of the wine’s acidity and sweetness.
* **Distribution:** Citric acid levels range from 0.00 to 0.56 g/dm³, with a mean value of 0.27 g/dm³. Higher citric acid levels are often associated with better quality wines due to their balanced taste.

#### Residual Sugar

* **Interpretation:** Residual sugar indicates the sweetness level of the wine. Moderate residual sugar can improve the wine’s flavor profile and balance its acidity.
* **Distribution:** This feature ranges from 1.0 to 15.5 g/dm³, with a mean value of 2.5 g/dm³. Wines with a higher residual sugar tend to receive higher quality ratings, as long as it is balanced with other attributes.

#### Chlorides

* **Interpretation:** Chlorides are related to salt content, which can influence the taste and preservation of the wine. Excessive chlorides can negatively affect the wine’s quality.
* **Distribution:** Chloride levels vary from 0.012 to 0.611 g/dm³, with a mean value of 0.045 g/dm³. Low chloride levels are preferable for higher quality wines.

#### Free Sulfur Dioxide and Total Sulfur Dioxide

* **Interpretation:** Sulfur dioxide acts as a preservative and antioxidant. High levels can protect the wine from oxidation and spoilage but can also impact the taste if excessive.
* **Distribution:** Free sulfur dioxide ranges from 1 to 72 mg/dm³, and total sulfur dioxide ranges from 6 to 289 mg/dm³. Generally, wines with balanced sulfur dioxide levels are rated higher for quality.

#### Density

* **Interpretation:** Density is related to the wine’s alcohol content and sugar levels. Higher density often indicates higher sugar or alcohol content.
* **Distribution:** Density ranges from 0.990 to 1.003 g/cm³, with a mean value of 0.996 g/cm³. Higher density wines can be associated with higher quality if the balance between sweetness and alcohol is well-maintained.

#### pH

* **Interpretation:** The pH level measures the wine’s acidity. Lower pH values indicate higher acidity, which can affect the wine’s taste and preservation.
* **Distribution:** pH levels range from 2.74 to 4.01, with a mean value of 3.30. Wines with balanced pH levels are typically rated higher in quality.

#### Sulphates

* **Interpretation:** Sulphates contribute to the wine’s stability and taste. They can enhance the wine’s flavor profile but excessive levels can impact quality.
* **Distribution:** Sulphates range from 0.33 to 2.0 g/dm³, with a mean value of 0.66 g/dm³. Optimal sulphate levels are associated with higher quality ratings.

#### Alcohol

* **Interpretation:** Alcohol content affects the wine’s body, flavor, and mouthfeel. Higher alcohol content is generally associated with richer and more flavorful wines.
* **Distribution:** Alcohol levels range from 8.4% to 14.9%, with a mean value of 10.5%. Wines with higher alcohol content often receive better quality ratings.

### Data Interpretation

The analysis reveals that certain features have a significant impact on the quality of red wine:

* **High Alcohol Content:** Generally associated with higher quality ratings. Alcohol contributes to the richness and complexity of the wine.
* **Balanced Acidity and Sweetness:** Wines with moderate levels of volatile acidity and residual sugar tend to receive higher ratings. Excessive acidity or sweetness can negatively affect the quality.
* **Optimal Levels of Sulfur Dioxide and Chlorides:** Balanced levels of sulfur dioxide and low chlorides are preferable for maintaining wine quality.

## 4. EDA Concluding Remarks

### Insights

The exploratory data analysis (EDA) provided valuable insights into the relationships between physicochemical attributes and wine quality. Key observations include:

* **Alcohol Content:** Higher alcohol content generally correlates with higher quality ratings. This is consistent with the idea that higher alcohol levels can enhance the wine's body and flavor.
* **Acidity Levels:** Both fixed and volatile acidity show a negative correlation with quality. Excessive acidity can negatively impact the taste and overall perception of the wine.
* **Residual Sugar:** A moderate level of residual sugar appears to be associated with higher quality ratings. This suggests that a balanced sweetness can improve the wine's appeal.

### Data Quality

The analysis also highlighted some data quality issues, including:

* **Missing Values:** Although the dataset is relatively complete, a small number of missing values were identified in certain features. These were addressed using imputation methods to ensure a consistent dataset.
* **Outliers:** Some outliers were detected in features such as alcohol content and residual sugar. These outliers were reviewed and handled appropriately to avoid skewing the model's performance.

## 5. Pre-processing Pipeline

### Handling Missing Values

Missing values were addressed through imputation techniques. For numerical features, the mean value was used for imputation, as it provides a reasonable estimate without introducing significant bias. This approach ensured that the dataset remained intact and usable for model training.

### Feature Encoding

Since all features in this dataset are numerical, encoding was not necessary. However, normalization was applied to ensure that all features contribute equally to the model. This step is crucial for models that are sensitive to feature scaling.

### Normalization/Standardization

To standardize the features, z-score normalization was applied. This process transforms the data to have a mean of 0 and a standard deviation of 1, ensuring that each feature contributes equally to the model. Standardization is particularly important for distance-based models such as k-nearest neighbors.

### Outlier Detection

Outliers were identified using statistical methods and visual inspection. For features like alcohol content and residual sugar, outliers were treated with caution. Extreme values were capped or adjusted to reduce their impact on the model, ensuring a more robust performance.

## 6. Building Machine Learning Models

### Model Selection

Several machine learning models were evaluated for their performance in predicting wine quality. The models selected for this project include:

* **Linear Regression:** A straightforward model that predicts quality based on linear relationships between features.
* **Decision Trees:** A non-linear model that splits the data based on feature values to make predictions.
* **Random Forests:** An ensemble model that combines multiple decision trees to improve accuracy and robustness.
* **Gradient Boosting:** A boosting model that iteratively improves predictions by focusing on errors from previous models.

### Training and Validation

The dataset was split into training and testing sets using an 80-20 split ratio. Cross-validation was employed to ensure that the model's performance is consistent across different subsets of the data. This approach helps in assessing the model's generalizability and avoiding overfitting.

### Evaluation Metrics

The following metrics were used to evaluate the performance of the models:

* **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions.
* **Mean Squared Error (MSE):** Assesses the average squared difference between predicted and actual values.
* **Root Mean Squared Error (RMSE):** Provides a measure of the standard deviation of prediction errors.
* **R² Score:** Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

### Results

The performance of the models was compared based on the evaluation metrics. The Random Forest model emerged as the best-performing model, achieving the lowest MAE, MSE, and RMSE values, and the highest R² score. This suggests that Random Forests are particularly effective for this prediction task, benefiting from their ensemble approach and ability to handle complex relationships between features.

## 7. Concluding Remarks

### Summary

The project successfully demonstrated the capability of machine learning models to predict red wine quality based on physicochemical attributes. Through thorough data analysis, pre-processing, and model evaluation, the Random Forest model was identified as the most effective tool for predicting wine quality with high accuracy. The model's performance underscores the importance of selecting appropriate features and employing robust modeling techniques.

### Impact

The ability to predict wine quality has significant implications for the wine industry. Wineries can use predictive models to enhance their quality control processes, optimize production methods, and cater to consumer preferences more effectively. By leveraging machine learning, wineries can make data-driven decisions that improve their products and maintain high standards.

### Future Work

Future work could focus on further improving the model's accuracy by incorporating additional features, such as sensory attributes or vineyard-specific data. Exploring more advanced models or hybrid approaches could also provide deeper insights and more precise predictions. Expanding the dataset to include a wider range of wine samples and attributes could enhance the model's generalizability and applicability across different wine varieties.