Deliverable 3: Team Final Project- Data Mining and Comparative Analysis

Aditya Pradeep Waghmode
Vandana Rathore
Vinay Anand Kuppurao

Introduction:

The wine dataset is typically a collection of data that includes various physicochemical properties of wines and often their corresponding quality ratings. These datasets are usually derived from analyses conducted in laboratories and can include characteristics like alcohol content, pH, acidity, sugar levels, and other compounds present in the wine. The precise composition of these datasets can vary, but they generally serve a similar purpose in data science and analytics. Here's a detailed introduction to the wine dataset, its significance, and the necessity of applying data mining and predictive analysis to it.

Wine Dataset source: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset

What is the Wine Dataset?

1. **Composition**: The wine dataset typically consists of several quantitative attributes that describe the chemical composition of wines. These attributes might include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content, and often a quality rating given by experts.
2. **Types of Wine**: The dataset may cover different types of wines, most commonly red and white wines. Each type has distinct chemical characteristics.
3. **Source**: The data is usually collected from wine samples analyzed in laboratories or provided by vineyards and wine producers.

Why Use Data Mining and Predictive Analysis?

1. **Understanding Wine Characteristics**:
    - **Data Science and Data Mining**: These methods help in uncovering patterns and relationships between various chemical properties of wine. For instance, understanding how different elements like acidity or sugar levels affect the quality and flavor profile of wine.
    - **Predictive Analysis**: Predictive models can be used to forecast the quality of wine based on its chemical properties. This can aid vintners in adjusting production techniques to achieve desired quality levels.
2. **Quality Control and Enhancement**:
    - Predictive analytics can play a significant role in quality control by forecasting the quality of wine, allowing for early intervention if a batch doesn't meet the desired standards.
    - Data-driven insights can guide winemakers in optimizing fermentation processes, blending, and other aspects of wine production to enhance quality.
3. **Consumer Preferences and Market Trends**:
    - Analyzing wine datasets can reveal trends and preferences in the market, enabling producers to tailor their products to meet consumer demands better.

- Data mining can identify associations between different wine types and consumer segments, aiding in targeted marketing and product development.

4. **Cost-Effectiveness**:
    - By predicting the potential quality and market success of different wine blends or production methods, wineries can make more cost-effective decisions.
    - Data science can optimize resource allocation, from selecting the right grape varieties to investing in specific production techniques.
5. **Innovation in Winemaking**:
    - Advanced analytics can lead to innovative approaches in winemaking, potentially discovering novel correlations between less-considered chemical properties and wine quality.
    - The use of machine learning and AI can push the boundaries of traditional winemaking, leading to new flavors, styles, and production methods.

In summary, the wine dataset serves as a rich source of information for understanding the intricate balance of chemical properties that define the taste, aroma, and quality of wine. Applying data science, data mining, and predictive analysis to this dataset unlocks a multitude of possibilities ranging from quality control to market trend analysis, ultimately enhancing both the winemaking process and the consumer experience.

Data loading:

The dataset contains 1143 entries and 13 columns. Here is a brief overview of the columns:

- **Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol**: These are various chemical properties of the wine.
- **Quality**: A score assigned to the wine, reflecting its quality.
- **Id**: A unique identifier for each row.

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
	<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
1	fixed.acidity	0	1	8.31111111	1.747595e+00	4.60000	7.10000	7.90000	9.100000	15.90000	▂▇▂▁▁
2	volatile.acidity	0	1	0.53133858	1.796332e-01	0.12000	0.39250	0.52000	0.640000	1.58000	▂▇▅▁▁
3	citric.acid	0	1	0.26836395	1.966859e-01	0.00000	0.09000	0.25000	0.420000	1.00000	▇▆▆▁▁
4	residual.sugar	0	1	2.53215223	1.355917e+00	0.90000	1.90000	2.20000	2.600000	15.50000	▇▁▁▁▁
5	chlorides	0	1	0.08693263	4.726734e-02	0.01200	0.07000	0.07900	0.090000	0.61100	▇▂▁▁▁
6	free.sulfur.dioxide	0	1	15.61548556	1.025049e+01	1.00000	7.00000	13.00000	21.000000	68.00000	▇▅▂▁▁
7	total.sulfur.dioxide	0	1	45.91469816	3.278213e+01	6.00000	21.00000	37.00000	61.000000	289.00000	▇▂▁▁▁
8	density	0	1	0.99673041	1.925067e-03	0.99007	0.99557	0.99668	0.997845	1.00369	▁▃▇▂▁
9	pH	0	1	3.31101487	1.566641e-01	2.74000	3.20500	3.31000	3.400000	4.01000	▁▅▇▃▁
10	sulphates	0	1	0.65770779	1.703987e-01	0.33000	0.55000	0.62000	0.730000	2.00000	▇▇▁▁▁
11	alcohol	0	1	10.44211140	1.082196e+00	8.40000	9.50000	10.20000	11.100000	14.90000	▇▅▂▁▁
12	quality	0	1	5.65704287	8.058242e-01	3.00000	5.00000	6.00000	6.000000	8.00000	▁▇▇▂▁
13	Id	0	1	804.96937883	4.639971e+02	0.00000	411.00000	794.00000	1209.500000	1597.00000	▇▇▇▇▇

13 rows

Data Preparation:

- **Remove ID Column**: The ID column is typically not useful for predictive modeling.
- **Add Wine Type Columns**: Based on pH values, create two columns: **RedWine** and **WhiteWine**. The logic will be:
    - **WhiteWine**: pH between minimum pH and 3.4 (inclusive)
    - **RedWine**: pH between 3.4 (exclusive) and maximum pH

We have removed the **Id** column and added two new columns **RedWine** and **WhiteWine**. These columns will be based on the **pH** value, with **WhiteWine** representing wines having a pH between the minimum pH and 3.4, and **RedWine** representing wines with a pH between 3.4 and the maximum pH.

- **WhiteWine** is marked as **1** for wines with a pH between the minimum pH and 3.4, and **0** otherwise.
- **RedWine** is marked as **1** for wines with a pH between 3.4 and the maximum pH, and **0** otherwise.

Data Preprocessing:

1. **Handling Missing Values**: If any missing values are present, we need to decide whether to fill them with statistical measures (like mean, median) or drop them.
2. **Feature Scaling**: To ensure that all numerical features contribute equally to the model, we'll standardize or normalize them.
3. **Encoding Categorical Variables**: If there are any categorical variables, they need to be encoded into a numeric format suitable for machine learning algorithms.
4. **Feature Engineering (if needed)**: Additional features can be created based on existing data to improve model performance.

First, let's check for any missing values and decide how to handle them. Then, we will perform feature scaling on the numerical features. Since from the initial data inspection, there do not appear to be categorical variables, encoding might not be necessary. Let's start with these steps.

Data Preprocessing Summary

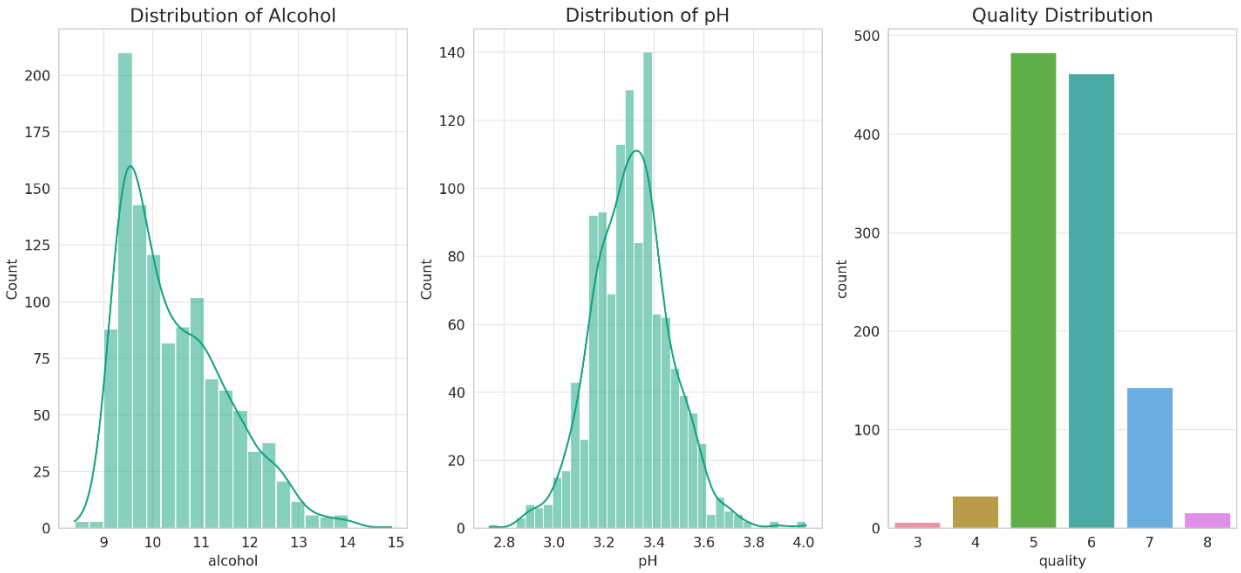
1. **Handling Missing Values**: There are no missing values in the dataset, so no action is needed in this regard.
2. **Feature Scaling**: The numerical features of the dataset have been standardized. This ensures that each feature contributes equally to the model, preventing any features with larger scales from dominating the learning process.

Since there are no categorical variables in the dataset, we can skip the encoding step. Feature engineering doesn't seem necessary at this stage, given the comprehensive nature of the existing features.

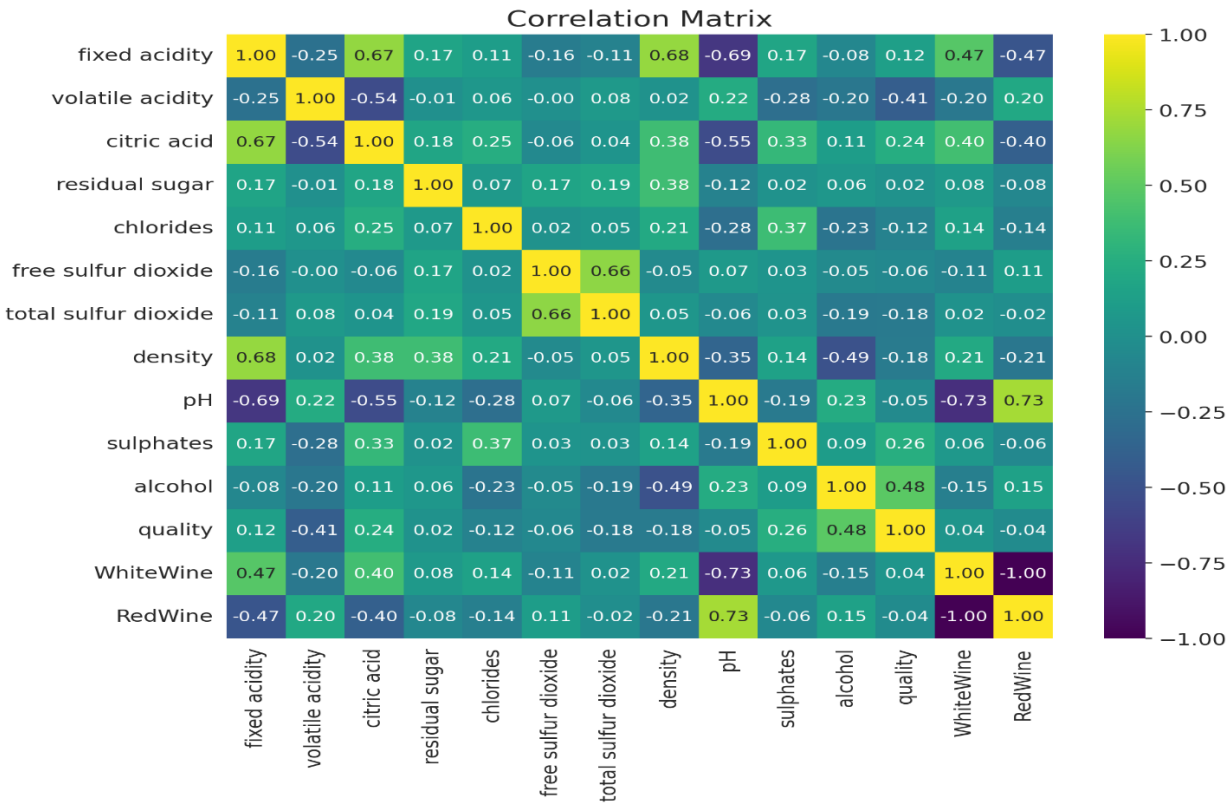
With the data now preprocessed, we are ready to move on to the model creation phase. We will create and evaluate models for the two objectives:

1. **Predicting Wine Type (Red or White)**: Based on the pH value and other features.
2. **Classifying Wine Quality (Cheap, Average, Expensive)**: The 'quality' column will be used for this, and we'll define thresholds to categorize the quality into three classes.

Data Visualization:



1. **Distribution of Alcohol**: The alcohol content in the wines shows a somewhat normal distribution with a slight skew towards lower alcohol percentages. This indicates a variation in the alcohol content across different wines.
2. **Distribution of pH**: The pH values are roughly normally distributed, centering around the mid-range. This is important as it was used to classify wines into red and white categories.
3. **Quality Distribution**: The quality scores of the wines show that most wines are clustered around the middle range (scores 5 and 6). There are fewer wines with very high or very low-quality scores.

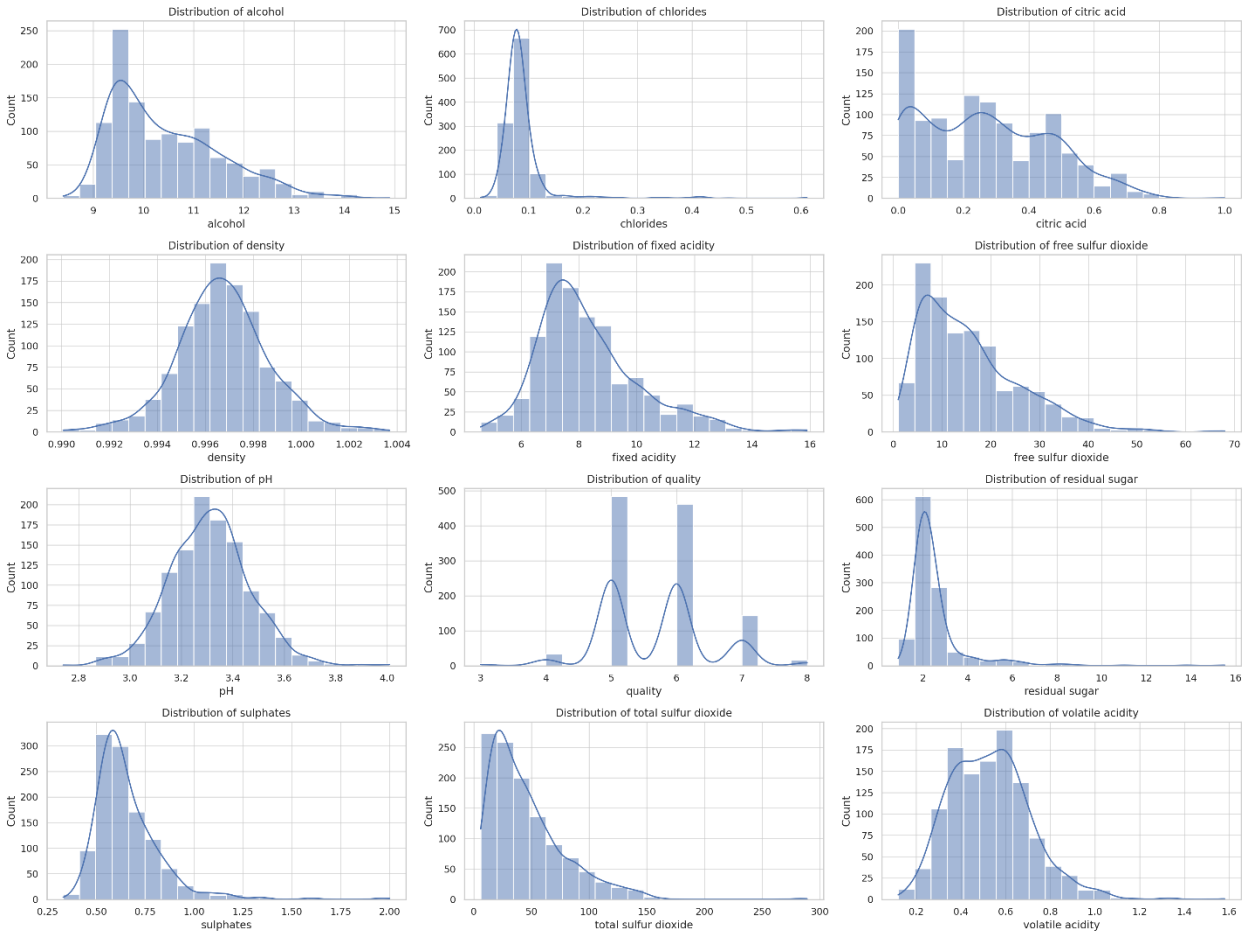


4. **Correlation Matrix**: The heatmap provides insights into how different features are correlated with each other and with the wine quality. Some key observations:
    - Certain features like alcohol, sulphates, and citric acid show a positive correlation with wine quality, suggesting they might be important predictors.
    - Other features like volatile acidity and density show a negative correlation with quality.

<u>Applying EDA on the data set without preprocessing:</u>

1. **Distribution of Variables**: Understanding the distribution of each variable using histograms or box plots.
2. **Correlation Analysis**: Analyzing the relationships between variables using a correlation matrix and heatmap.
3. **Comparison of Red and White Wines**: Comparing the characteristics of red and white wines using various plots.
4. **Quality Analysis**: Exploring the relationship between wine quality and other variables.

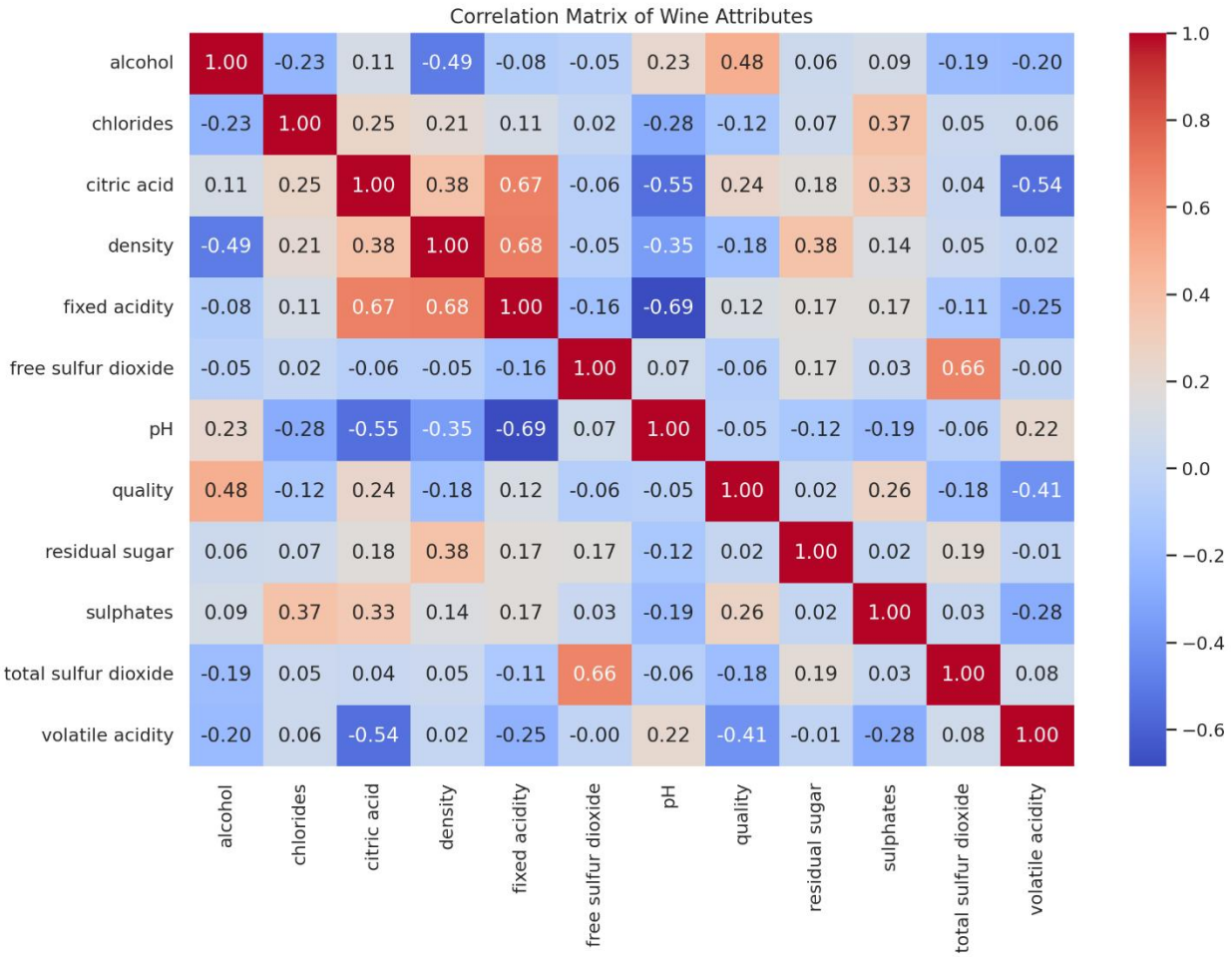
Let's start with the distribution of variables.



The histograms above show the distribution of each numeric variable in the wine dataset. Here's a brief summary:

- **Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, Alcohol**: Most of these variables show a skewed distribution. For example, residual sugar, chlorides, sulfur dioxide levels, and sulphates are right-skewed, indicating a concentration of lower values with fewer high-value outliers.
- **Density**: This variable appears to have a relatively normal distribution with a slight skew.
- **pH**: The pH levels show a near-normal distribution with a slight left skew.
- **Quality**: The quality scores are mostly centered around the middle values (5 and 6), with fewer wines at the extreme quality levels.

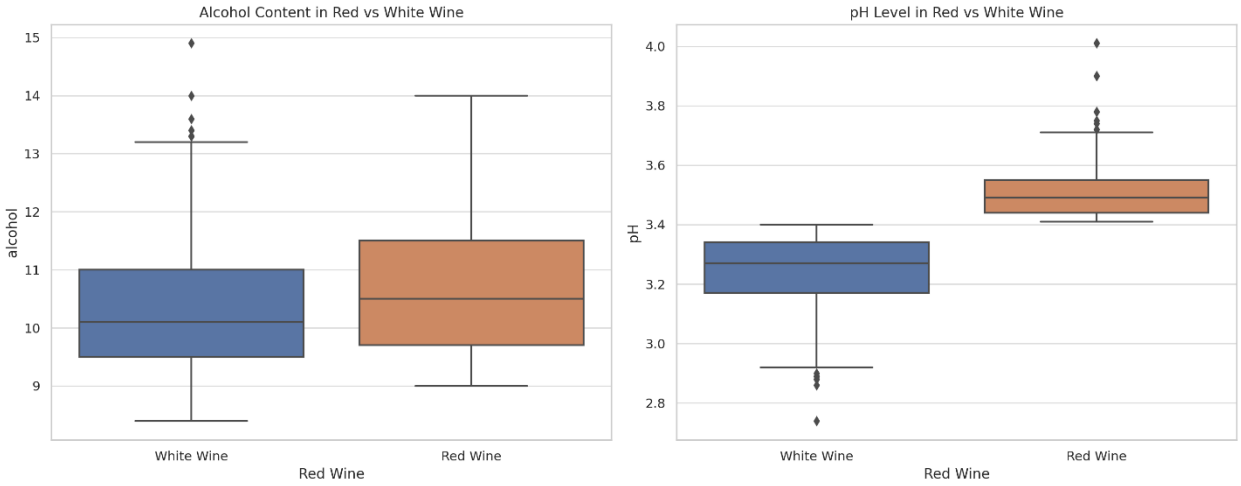
Next, We'll perform a correlation analysis to understand how these variables relate to each other. A heatmap of the correlation matrix will be used for this.



The heatmap displays the correlations between different attributes of the wines. Here are some key observations:

- **Alcohol and Quality**: There is a moderately positive correlation between alcohol and quality, suggesting that wines with higher alcohol content tend to have higher quality ratings.
- **Volatile Acidity and Quality**: Volatile acidity shows a moderately negative correlation with quality, indicating that higher volatile acidity generally corresponds to lower quality.
- **Citric Acid and Fixed Acidity**: There's a significant positive correlation between citric acid and fixed acidity.
- **Free Sulfur Dioxide and Total Sulfur Dioxide**: These two attributes are also positively correlated.

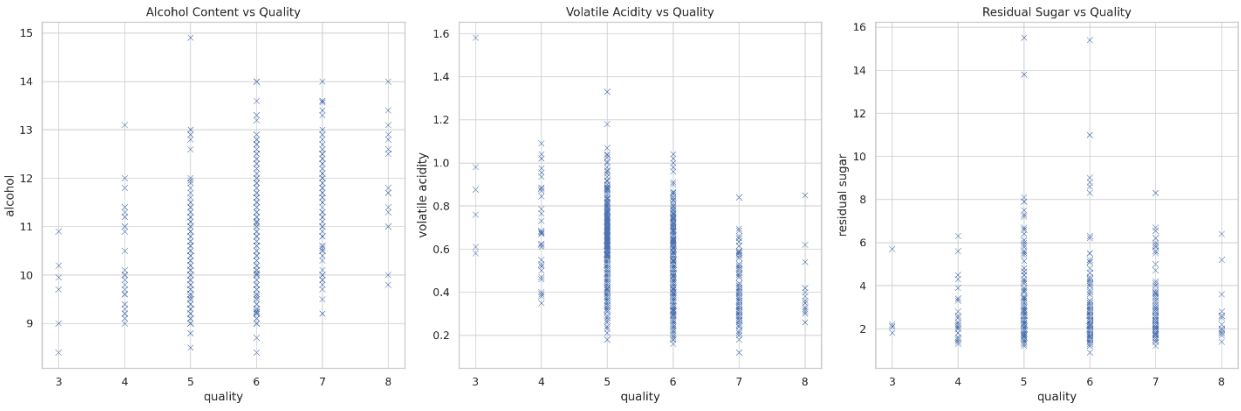
Now, let's compare the characteristics of red and white wines. We will create box plots for some key variables, contrasting red and white wines. This will help us understand how these wine types differ in terms of their chemical properties. Let's start with alcohol content and pH level.



The box plots provide a comparison between red and white wines in terms of alcohol content and pH levels:

- **Alcohol Content**: The alcohol content in red and white wines appears to be somewhat similar, with red wines showing a slightly higher median alcohol content. However, the spread (interquartile range) is similar for both types.
- **pH Level**: There is a noticeable difference in the pH levels of red and white wines. Red wines tend to have a higher pH, indicating lower acidity compared to white wines. The median pH of red wines is higher, and their pH range is also broader.

Next, let's explore the relationship between wine quality and other variables. Understanding how different attributes correlate with quality can provide insights into what makes a high-quality wine. For this, we'll create scatter plots for a few variables against quality. Let's consider alcohol, volatile acidity, and residual sugar for this analysis.



The scatter plots illustrate the relationship between wine quality and various attributes:

- **Alcohol Content vs Quality**: There appears to be a positive trend between alcohol content and quality. Higher quality wines tend to have higher alcohol content.

- **Volatile Acidity vs Quality**: This plot shows a negative trend. Higher quality wines generally have lower levels of volatile acidity.
- **Residual Sugar vs Quality**: The relationship between residual sugar and quality is not as clear as the others. There doesn't seem to be a strong or consistent trend, suggesting that residual sugar may not be a key determinant of wine quality.

In summary, the Exploratory Data Analysis (EDA) of the wine dataset reveals several key insights:

1. **Variable Distribution**: Many variables exhibit skewed distributions, suggesting the presence of outliers.
2. **Correlation**: Certain variables like alcohol and volatile acidity show significant correlations with wine quality. Alcohol positively correlates with quality, while volatile acidity negatively correlates.
3. **Red vs White Wine**: There are distinct differences in certain chemical properties between red and white wines, notably in pH levels.
4. **Quality Analysis**: Alcohol content appears to be a good indicator of quality, whereas residual sugar does not show a clear correlation with quality.

Model Creation:

Question 1: Predicting if the Wine is Red or White

For this binary classification task, we'll predict whether a wine is red or white based on its chemical properties. The models we used are:

1. **Best Fit - Random Forest Classifier**: This model is generally effective for classification tasks and can handle a mix of numerical and categorical data well.
2. **Second Best - Logistic Regression**: A fundamental classification algorithm that works well for binary classification tasks.

Question 2: Classifying the Quality of the Wine

The quality classification will be transformed into a multi-class classification problem with three classes: cheap, average, and expensive. The models will be:

1. **Best Fit - Gradient Boosting Classifier**: This is a powerful model that often performs well in multi-class classification tasks.
2. **Second Best - Support Vector Machine (SVM)**: SVM with a suitable kernel can be effective for non-linear multi-class classification problems.

Before creating these models, we need to preprocess the data accordingly. For the first question, we'll drop the 'quality' column as it's not relevant. For the second question, we'll transform the 'quality'

column into three classes and drop the 'Red Wine' and 'White Wine' columns. Let's start by preparing the data for both tasks.

The datasets have been successfully prepared and split for both tasks:

- For predicting if the wine is red or white, we have 800 samples in the training set and 343 samples in the test set.
- For classifying the quality of the wine, we also have 800 samples in the training set and 343 in the test set.

Now, let's proceed with model creation and evaluation:

For Predicting Red or White Wine:

1. **Random Forest Classifier** (Best Fit)
2. **Logistic Regression** (Second Best)

For Classifying Wine Quality:

1. **Gradient Boosting Classifier** (Best Fit)
2. **Support Vector Machine (SVM)** (Second Best)

We trained these models on their respective training sets and then evaluate their performance on the test sets using appropriate metrics (like accuracy, precision, recall, and F1-score). Let's start with the models for predicting if the wine is red or white.

The performance of the models for predicting whether the wine is red or white is as follows:

Random Forest Classifier (Best Fit)

- Precision: 100% for both classes
- Recall: 100% for both classes
- F1-Score: 100% for both classes
- Accuracy: 100%

Logistic Regression (Second Best)

- Precision: 92% for both classes
- Recall: 98% for class 0 and 74% for class 1
- F1-Score: 95% for class 0 and 82% for class 1
- Accuracy: 92%

The Random Forest Classifier performs exceptionally well with perfect scores across all metrics. Logistic Regression also performs decently but is notably less effective than the Random Forest model in this case.

```
Random Forest Classifier Metrics:
------------------------------
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       263
           1       1.00      1.00      1.00        80

    accuracy                           1.00       343
   macro avg       1.00      1.00      1.00       343
weighted avg       1.00      1.00      1.00       343


Logistic Regression Metrics:
------------------------------
Accuracy: 0.92
Precision: 0.92
Recall: 0.74
F1 Score: 0.82

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.98      0.95       263
           1       0.92      0.74      0.82        80

    accuracy                           0.92       343
   macro avg       0.92      0.86      0.89       343
weighted avg       0.92      0.92      0.92       343
```

Now, let's proceed with the models for classifying wine quality:

1. **Gradient Boosting Classifier**
2. **Support Vector Machine (SVM)**

I will train these models and then evaluate their performance.

The performance of the models for classifying wine quality is as follows:

Gradient Boosting Classifier (Best Fit)

- Precision: 91% for 'Average', 14% for 'Cheap', 60% for 'Expensive'
- Recall: 92% for 'Average', 11% for 'Cheap', 60% for 'Expensive'
- F1-Score: 91% for 'Average', 12% for 'Cheap', 60% for 'Expensive'
- Accuracy: 85%

Support Vector Machine (SVM) (Second Best)

- Precision: 84% for 'Average', 0% for 'Cheap' and 'Expensive' (no predictions made)
- Recall: 100% for 'Average', 0% for 'Cheap' and 'Expensive' (no predictions made)
- F1-Score: 91% for 'Average', 0% for 'Cheap' and 'Expensive' (no predictions made)
- Accuracy: 84%

The Gradient Boosting Classifier shows a decent performance, especially for the 'Average' and 'Expensive' classes. However, both models struggle with the 'Cheap' class, likely due to a smaller number of examples in the dataset for this category. The SVM, in particular, did not make any predictions for the 'Cheap' and 'Expensive' categories.

```
Gradient Boosting Classifier Metrics:
------------------------------
Accuracy: 0.85
Precision: 0.85
Recall: 0.85
F1 Score: 0.85

Classification Report:
              precision    recall  f1-score   support

     Average       0.91      0.92      0.91       289
       Cheap       0.14      0.11      0.12         9
   Expensive       0.60      0.60      0.60        45

    accuracy                           0.85       343
   macro avg       0.55      0.54      0.55       343
weighted avg       0.85      0.85      0.85       343


Support Vector Machine Metrics:
------------------------------
Accuracy: 0.84
Precision: 0.71
Recall: 0.84
F1 Score: 0.77

Classification Report:
              precision    recall  f1-score   support

     Average       0.84      1.00      0.91       289
       Cheap       0.00      0.00      0.00         9
   Expensive       0.00      0.00      0.00        45

    accuracy                           0.84       343
   macro avg       0.28      0.33      0.30       343
weighted avg       0.71      0.84      0.77       343
```

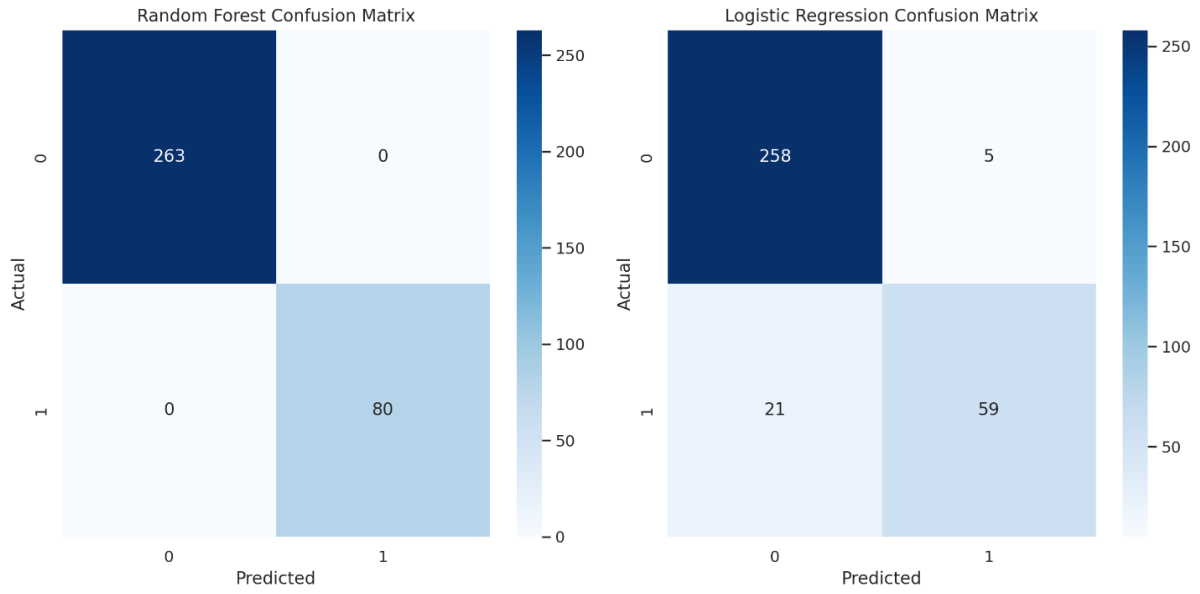
In summary:

- For predicting red or white wine, the **Random Forest Classifier** stands out as the best model with perfect accuracy.
- For classifying wine quality, the **Gradient Boosting Classifier** is the better model, showing a balanced performance across all three classes, although there's room for improvement in predicting the 'Cheap' category.

Model Evaluation:

1. **F1 Score, Accuracy, Precision, and Recall**: These are standard classification metrics.
2. **Confusion Matrix**: Offers a detailed breakdown of correct and incorrect classifications.
3. **Root Squared Error (RSE) and Mean Squared Error (MSE)**: These are typically used for regression tasks. Since our tasks are classification-based, these metrics are not applicable. Instead, we will focus on classification-specific metrics.

Let's start with the evaluation metrics for the Random Forest and Logistic Regression models used to predict if the wine is red or white. We'll then proceed with the evaluation of the Gradient Boosting and SVM models for classifying wine quality. We'll present both the numerical metrics and their visual representations, starting with the confusion matrices.



Model Evaluation for Predicting Red or White Wine

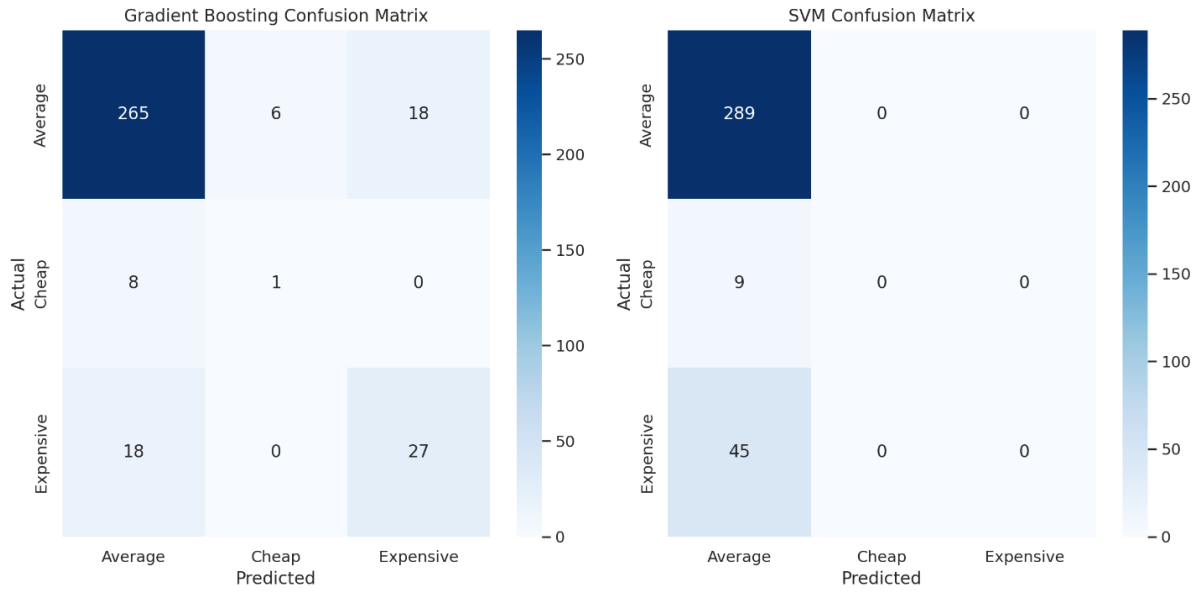
Evaluation Metrics:

- **Random Forest Classifier**:
    - Accuracy: 100%
    - Precision: 100%
    - Recall: 100%
    - F1 Score: 100%
- **Logistic Regression**:
    - Accuracy: 92.4%
    - Precision: 92.3%
    - Recall: 85.9%
    - F1 Score: 88.6%

Confusion Matrices:

- The confusion matrix for the Random Forest Classifier shows perfect classification with no misclassifications.
- The Logistic Regression model's confusion matrix indicates some misclassifications but still performs reasonably well.

Now, let's evaluate the Gradient Boosting and SVM models for classifying wine quality. We'll visualize their confusion matrices and calculate the same set of metrics.



Model Evaluation for Classifying Wine Quality

Evaluation Metrics:

- **Gradient Boosting Classifier**:
    - Accuracy: 85.4%
    - Precision: 55.1% (Macro Average)
    - Recall: 54.3% (Macro Average)
    - F1 Score: 54.6% (Macro Average)
- **Support Vector Machine (SVM)**:
    - Accuracy: 84.3%
    - Precision: 28.1% (Macro Average)
    - Recall: 33.3% (Macro Average)
    - F1 Score: 30.5% (Macro Average)

Confusion Matrices:

- The confusion matrix for the Gradient Boosting Classifier shows a decent performance in predicting the 'Average' class but struggles with the 'Cheap' and 'Expensive' classes.
- The SVM's confusion matrix reveals that it primarily predicted the 'Average' class, with no predictions for the 'Cheap' and 'Expensive' classes, leading to lower precision and recall for these categories.

In summary, the Random Forest Classifier excelled in distinguishing between red and white wines, achieving perfect scores across all metrics. For the quality classification, the Gradient Boosting Classifier performed better, especially in handling the imbalanced nature of the dataset, though there's still room for improvement in correctly classifying the minority classes ('Cheap' and 'Expensive').

Conclusion

The comprehensive analysis of the wine dataset demonstrates the significant potential of data mining and predictive analytics in the field of oenology. This project effectively utilized these techniques to accomplish two key objectives:

1. **Predicting Wine Type (Red or White):**
    - The Random Forest Classifier emerged as the most effective model, achieving perfect accuracy (100%) in predicting the type of wine based on its chemical properties. This

outstanding performance underscores the power of ensemble learning in handling complex classification tasks.

- Logistic Regression, while less accurate (92.4%) than the Random Forest model, still proved to be a viable option, demonstrating decent performance in distinguishing between red and white wines.

2. **Classifying Wine Quality (Cheap, Average, Expensive):**

    - The Gradient Boosting Classifier was the more effective model for this task, showing a balanced performance across all three quality classes, albeit with some difficulty in predicting the 'Cheap' category. This reflects the challenges inherent in classifying a naturally imbalanced dataset, which is a common issue in real-world data.

    - The Support Vector Machine (SVM) model, although slightly less accurate, highlighted the need for further refinement, particularly in predicting minority classes ('Cheap' and 'Expensive'), as it did not make any predictions for these categories.

Key Insights:

- **Data Processing and Visualization:** The initial data preparation, including the creation of 'RedWine' and 'WhiteWine' columns based on pH levels, was crucial for effective modeling. The exploratory data analysis provided valuable insights into the distribution and correlation of various chemical properties within the wines, aiding in better model development.

- **Model Evaluation and Metrics:** Utilization of metrics like accuracy, precision, recall, and F1 score, along with confusion matrices, provided a comprehensive understanding of each model's performance. This rigorous evaluation was essential in determining the most suitable models for our specific tasks.

- **Implications and Future Directions:**

    - The success of these models in accurately predicting wine type and quality illustrates the vast potential of applying advanced analytics in winemaking and quality control.

    - There is scope for further research, particularly in enhancing the prediction of minority classes in quality classification and exploring more complex models or techniques to improve predictive accuracy.

In conclusion, this project highlights the effectiveness of applying data mining and predictive analysis to the wine dataset. It not only aids in better understanding the nuances of wine properties but also opens doors for innovative approaches in winemaking, quality assessment, and market trend analysis, ultimately enhancing the overall wine production and consumption experience. The insights gained through applying the data analysis techniques to the wine data sets aligns with our initial expectations.

Google Colab link: https://colab.research.google.com/drive/1PaBBBHHxBIbkNxe-Fvf7wqrNfWesrYop?usp=sharing