Intermediate Project Report Sample: Wine Quality Analysis

1 Introduction & Objective

The global wine market has seen substantial growth over the past decade, making quality assessment an essential part of winemaking. This project examines the influence of chemical attributes on wine quality and aims to determine if these attributes can also help distinguish red wines from white wines. The main objectives of this project are to:

- Identify the key chemical predictors (e.g., alcohol content, volatile acidity, sulphates) that significantly affect wine quality.
- Explore differences in quality determinants between red and white wines.
- Evaluate basic modeling techniques for regression (predicting quality scores) and classification (distinguishing wine types).

2 Data and Exploratory Data Analysis (EDA)

The dataset used in this study includes several chemical measurements such as fixed and volatile acidity, citric acid, residual sugar, chlorides, free and total sulfur dioxide, pH, density, sulphates, and alcohol. Additionally, the dataset contains sensory quality scores (rated on a scale from 1 to 10) and a classification variable indicating the type of wine (red or white).

The initial EDA involved:

- **Data Cleaning:** Verifying the dataset for missing values and outliers. If any were identified, appropriate data transformations were applied to adjust variables measured on different scales.
- **Statistical Summary:** Calculating key statistics (mean, median, standard deviation) to understand the distribution of predictors for both red and white wines.

- **Visual Analysis:** Creating scatter plots, box plots, and correlation matrices to identify trends and potential relationships between chemical attributes and wine quality.

Figure 1 shows an example scatter plot that illustrates the relationship between alcohol content and wine quality.

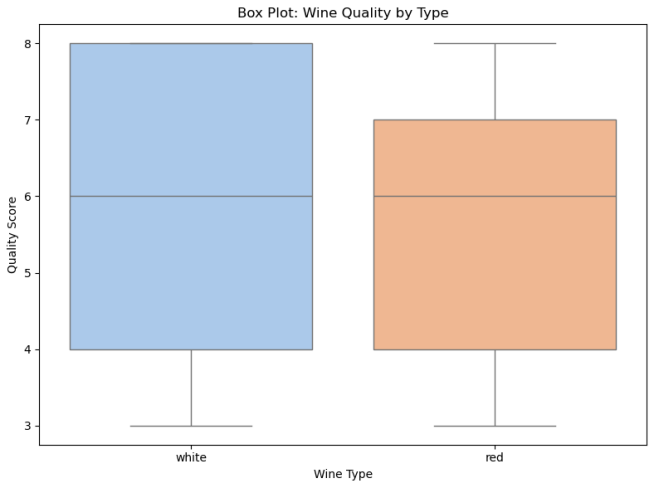

Figure 1: Example scatter plot showing the relationship between alcohol content and wine quality.

Preliminary results suggested that alcohol content and volatile acidity are consistently important predictors. For instance, higher alcohol levels generally correspond to better quality scores, while increased volatile acidity tends to reduce quality.

3 Modeling Approaches

To address the research questions, several basic modeling techniques were applied.

3.1 Regression Analysis

For predicting wine quality, linear regression models were developed using different variable selection methods. The following approaches were explored:

- **Full Model and Stepwise Selection:** Multiple linear regression was performed using the full set of predictors as well as forward and backward selection methods. Although the performance across these models was similar, alcohol and volatile acidity consistently emerged as key predictors.
- **Regularization:** Basic regularization methods such as Ridge and Lasso regression were applied to reduce multicollinearity. These methods helped to simplify the models by shrinking the coefficients of less important variables.

Table 1 shows example test MSE values for red and white wines using selected regression methods.

Method	Red Wine Test MSE	White Wine Test MSE
Full Model	0.4341	0.5558
Forward Selection	0.4341	0.5558
Backward Selection	0.4354	0.5554
Lasso Regression	0.4381	0.5703
Ridge Regression	0.4355	0.5697

Table 1: Test MSE for various regression models. (Placeholder values)

3.2 Classification Analysis

To determine whether chemical attributes could be used to classify wines as red or white, several classification methods were implemented:

- **Logistic Regression and Linear Discriminant Analysis (LDA):** Both methods were used to model the probability of a wine being red or white. They achieved very low error rates and effectively identified the influence of key predictors such as fixed acidity and volatile acidity.
- **Basic Decision Tree:** A simple decision tree classifier was also applied to provide an interpretable model. Although its error rate was slightly higher than that of logistic regression and LDA, it served as a useful benchmark.

Table 2 presents example test error rates for these classification techniques.

Method	Test Error Rate
Logistic Regression	0.0074
Linear Discriminant Analysis (LDA)	0.0062
Decision Tree	0.0209

Table 2: Test error rates for classification models. (Placeholder values)