MTH 786P Wine Quality Prediction

Prapthi Harish
230028139

1 Problem Statement

Implement, describe and present regression and/or classification models of your choice to predict the quality of white wines given a range of their features. The goal is to develop regression and/or classification models using any number of the variables provided, which describe wines' features, to predict their quality(measured as a score from 0 to 10 based on sensory data from three experts).

1.1 Introduction

The dataset includes features such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Each row in the dataset represents a sample of white wine, with its corresponding quality score. To address this problem, we will explore regression algorithms like linear regression, lasso regression,ridge regression and classification algorithm like logistic regression, Support Vector Machine algorithms.To evaluate the performance of the model we use the metrics like Root Mean Squared Error (RMSE) and Mean Squared Error(MSE) for regression and accuracy, precision, recall, and F1 score for classification. [4]

2 Analysis

Checking for Missing Values: Missing values can compromise data integrity. Ensuring a complete dataset involves identifying and addressing any missing values.

```
# missing values
print(df.isnull().sum())
```

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

Figure 1: Checking the Missing Values

Summary of Statistics: Quickly analyzing central tendencies and spreads of numerical features aids in understanding the dataset.

```
In [6]: summary_stats = df[Float].describe()
        print(summary_stats)
```

```
       fixed acidity  volatile acidity  citric acid  residual sugar  \
count    4898.000000       4898.000000  4898.000000     4898.000000
mean        6.854788          0.278241     0.334192        6.391415
std         0.843868          0.100795     0.121020        5.072058
min         3.800000          0.080000     0.000000        0.600000
25%         6.300000          0.210000     0.270000        1.700000
50%         6.800000          0.260000     0.320000        5.200000
75%         7.300000          0.320000     0.390000        9.900000
max        14.200000          1.100000     1.660000       65.800000

         chlorides  free sulfur dioxide  total sulfur dioxide      density  \
count  4898.000000          4898.000000           4898.000000  4898.000000
mean      0.045772            35.308085            138.360657     0.994027
std       0.021848            17.007137             42.498065     0.002991
min       0.009000             2.000000              9.000000     0.987110
25%       0.036000            23.000000            108.000000     0.991723
50%       0.043000            34.000000            134.000000     0.993740
75%       0.050000            46.000000            167.000000     0.996100
max       0.346000           289.000000            440.000000     1.038980

                pH    sulphates      alcohol
count  4898.000000  4898.000000  4898.000000
mean      3.188267     0.489847    10.514267
std       0.151001     0.114126     1.230621
min       2.720000     0.220000     8.000000
25%       3.090000     0.410000     9.500000
50%       3.180000     0.470000    10.400000
75%       3.280000     0.550000    11.400000
max       3.820000     1.080000    14.200000
```

Figure 2: Summary of Statistics

Frequency Distribution: Understanding the frequency of each unique value or range is crucial for insights into data shape, distribution, and central tendency. This facilitates informed decision-making and helps identify outliers and patterns.

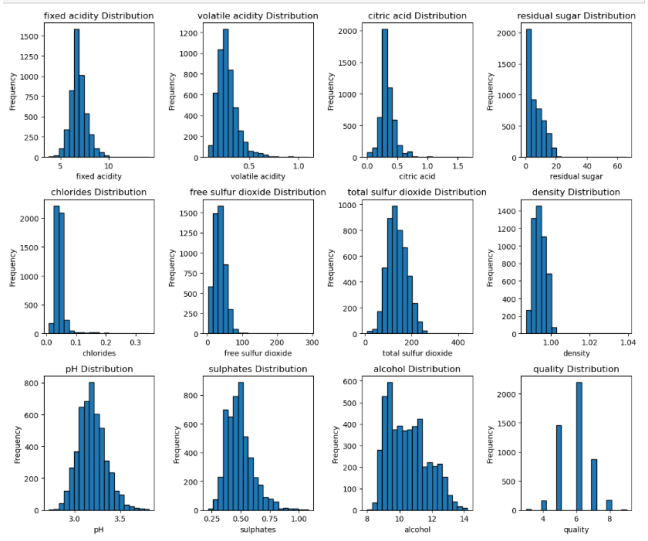


Figure 3: Frequency Distribution of Different Variables

1. Fixed Acidity Distribution: Concentrated around 6 to 8 grams per deciliter, right-skewed.

2. Volatile Acidity Distribution: Heavily skewed to the right, clustering at lower values.

3. Citric Acid Distribution: Right-skewed, concentrated close to 0.

4. Residual Sugar Distribution: Extremely right-skewed, most wines with low sugar.

5. Chlorides Distribution: Right-skewed, bulk at lower concentrations, with a tail of higher values.

6. Free Sulfur Dioxide Distribution: Right-skewed, peak at low levels, a tail of very high levels.

7. Total Sulfur Dioxide Distribution: Right-skewed, most wines with lower levels, a tail of high values.

8. Density Distribution: Appears normally distributed, centered around 0.996.

9. pH Distribution: Roughly normal, slight left skew, concentrated between 3.0 and 3.5.

10. Sulphates Distribution: Right-skewed, concentrated at lower levels, tail towards higher values.

11. Alcohol Distribution: Roughly normal, slight right skew, more wines towards the center.

12. Quality Distribution: Concentrated around 5 to 6, fewer at extremes, slight left skew.

Data Distribution: Box plots are useful for visualizing the central tendency, dispersion, and skewness of data distributions, as well as identifying outliers.

1. Fixed Acidity Box Plot: The distribution has a median value around 7, with most data points falling between approximately 6 and 8. There are several outliers on both the lower and higher ends of the scale, indicating some wines with very low or high fixed acidity.

2. Volatile Acidity Box Plot: This box plot shows a tight distribution with a median near 0.3. The interquartile range (IQR) is small, indicating less variability in volatile acidity. There are many outliers present, suggesting some wines have significantly higher volatile acidity.

3. Citric Acid Box Plot: The median citric acid content is around 0.3. The IQR is relatively compact, suggesting that citric acid levels are consistent across many wines. Several outliers indicate that a few wines have unusually high citric acid.

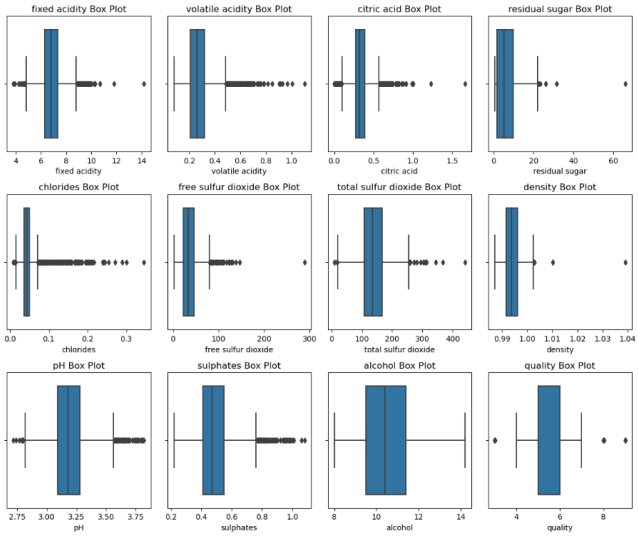


Figure 4: Data Distribution of Different Variables

4. Residual Sugar Box Plot: Here, the median is close to 2, but the IQR spans from about 1 to 3. There are numerous outliers, especially on the higher end, indicating that some wines have much higher residual sugar content.

5. Chlorides Box Plot: The median chlorides level is just above 0.05, with a narrow IQR. There are several outliers on the higher end, indicating a few wines with exceptionally high chloride levels.

6. Free Sulfur Dioxide Box Plot: The median is around 30, with a wide IQR stretching from approximately 15 to 45, indicating variability in free sulfur dioxide levels among wines. There are outliers on the higher end, showing that some wines have very high free sulfur dioxide content.

7. Total Sulfur Dioxide Box Plot: This plot has a median value close to 120, with an IQR from around 80 to 160. There are several outliers indicating some wines with very high total sulfur dioxide levels.

8. Density Box Plot: The median density is slightly above 0.996, with a very narrow IQR, suggesting that most wines have a similar density. A few outliers indicate some variability, with a few wines having notably higher or lower density.

9. pH Box Plot: This plot has a median pH value around 3.3, with an IQR from about 3.1 to 3.5. The distribution has a few outliers on either end, indicating some wines with particularly high or low pH levels.

10. Sulphates Box Plot: The median is around 0.5, with a relatively tight IQR. There are outliers, especially on the higher end, suggesting some wines contain much higher sulphate levels.

11. Alcohol Box Plot: The median alcohol content is around 10.5

12. Quality Box Plot: The quality scores have a median around 6, with an IQR from 5 to 6.5. There are outliers on both the lower and higher ends, which shows that there are wines of varying quality outside the normal range.

Correlation Coefficients: Each cell in the matrix shows the correlation between two variables, indicated by the row and column headers.

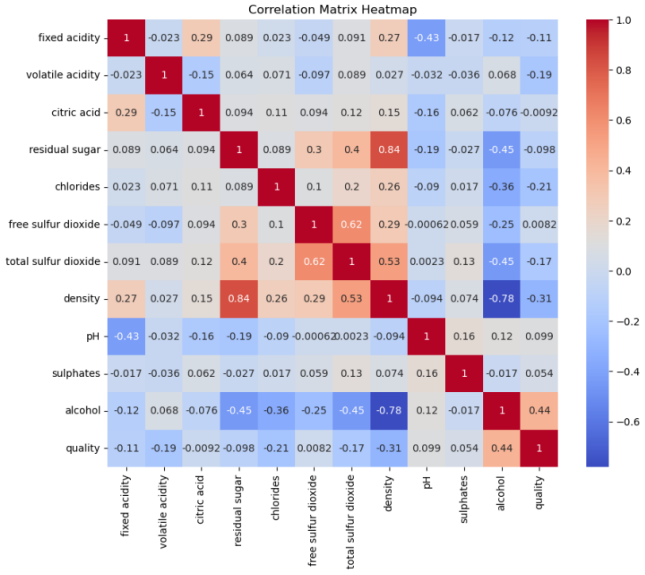


Figure 5: Correlation Matrix Heatmap

1. Alcohol and Quality: The relationship between alcohol content and wine quality is moderately strong and positive (0.44), implying that wines with higher alcohol levels are likely to receive better quality ratings.

2. Density and Alcohol: A significant negative correlation (-0.78) exists between density and alcohol content, indicating that wines with higher alcohol levels tend to have lower density.

3. pH and Fixed Acidity: The pH level and fixed acidity exhibit a moderately strong negative correlation (-0.43), aligning with the expectation that acidity and pH are inversely related.

4. Citric Acid and Fixed Acidity: Citric acid and fixed acidity show a moderate negative correlation (-0.29), suggesting that wines with higher fixed acidity may have lower citric acid levels.

5. Total Sulfur Dioxide and Free Sulfur Dioxide: Total sulfur dioxide and free sulfur dioxide display a strong positive correlation (0.62), indicating that wines with higher free sulfur dioxide levels also tend to have higher total sulfur dioxide content.

6. Density and Residual Sugar: A very strong positive correlation (0.84) exists between density and residual sugar, implying that wines with higher sugar content are denser.

7. Quality and Other Variables: Quality demonstrates no strong correlations with most other variables. However, there are slight negative correlations with volatile acidity (-0.19), density (-0.31), and chlorides (-0.21), suggesting that higher quality wines might have lower values in these properties.

3 Methods

- **Linear Regression**:Linear regression assumes a linear relationship between input features x and output y.Calculates the weights using the normal equation and predicts the target variable for the test set.

$$f(\mathbf{x_i}) = w_0 + \sum_{j=1}^{d} w_j x_{ij}$$

Linear transformation of vector $\mathbf{x_i} = (x_{i1}, \ldots, x_{id})$ with weights $\mathbf{w} \in \mathbb{R}^{d+1}$

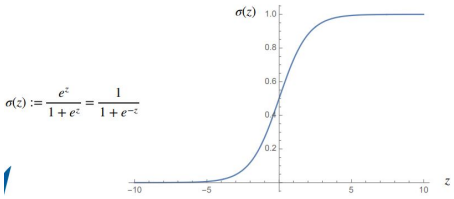
- **Lasso Regression** [5]:The lasso regression function predicts the target variable for the test set by iteratively updating the weights for each feature through the use of coordinate descent.It increases the cost function's penalty term. The coefficients' absolute sum is represented by this term. This term penalises, causing the model to decrease the value of coefficients in order to reduce loss, as the coefficients' value increases from 0.

$$\mathbf{w}_\alpha = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \alpha \|\mathbf{w}\|_1 \right\}$$

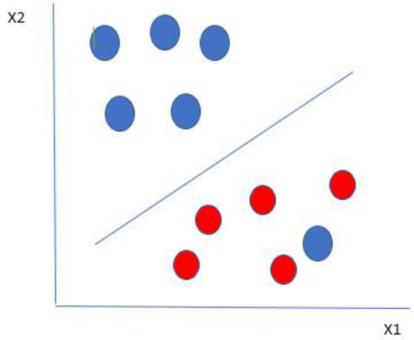
- **Ridge Regression** [3] :A penalty term equal to the square of the coefficient is added in ridge regression. The square of the coefficients' magnitudes represents the L2 term.Lasso regression tends to set coefficient values to zero, whereas Ridge regression never does. This is the difference between the two regression methods.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}$$

- **Logistic Regression** [2]:The logistic regression algorithm, which is mainly used for binary classification, makes use of the logistic function, also known as a sigmoid function. It generates a probability value between 0 and 1 based on inputs that are entered as independent variables. With the exception of how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification problems, linear regression is used to solve regression problems.



- **Support Vector Machine** [1]:The goal of the SVM algorithm is to locate the best hyperplane in an N-dimensional space that will allow the feature space's data points belonging to various classes to be separated.



Hyperplane which is the most optimized one

4 Results

Model	RMSE	MSE
Linear Regression	0.7122	0.50721
Logistic Regression	0.5459	0.29796
Support Vector Machine	0.9476	0.89796
Lasso Regression	0.7148	0.51091
Ridge Regression	0.7148	0.51089

Table 1: Summary of Model Performance

- **Linear and Lasso Regression**:Linear Regression and Lasso Regression show similar performance, with slightly higher errors compared to Logistic Regression.
- **Ridge Regression**:Ridge Regression introduces regularization to handle potential issues with multicollinearity. It performs decently but has slightly higher errors compared to Logistic Regression.
- **Logistic Regression**:Logistic Regression, often used for binary classification,providing the lowest errors among the models.However, Logistic regression is better suited for classification tasks rather than a continuous regression problem like predicting wine quality scores.
- **Support Vector Machine**:SVM, a powerful classification algorithm, shows the highest errors in this case. It might not be the best choice for this dataset or may require further optimization.

The findings emphasise how crucial it is to test out different models and comprehend the properties of the dataset in order to select the best algorithm for predictive modeling.

5 Conclusion

Regression models aim to predict a continuous target variable (in this case, the quality score), while classification models categorize the samples into discrete classes or quality levels.The wine quality variable appears to follow a linear-style trend with the physicochemical features provided in the data based on intuition and domain knowledge. By treating it as a classification, Logistic Regression is able to predict the discrete quality rating classes better than the linear regression model. However, given that the goal is to predict the quality of white wines, which is a regression task (since the quality is a continuous variable ranging from 0 to 10) hence, Linear Regression is the most suitable model for this problem statement.Using linear regression over logistic regression since preserving the numerical scores and linear relationships will allow better modeling of the quality scores.

References

[1] Manisha Koranga a, Richa Pandey a, Mayurika Joshi a, and Manish Kumar. Analysis of white wine using machine learning algorithms. 2020.

[2] Klaus Backhaus, Bernd Erichson, Sonja Gensler, and Rolf Weiber. Logistic regression. 2022.

[3] K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire. Prediction of wine quality using machine learning. 2021.

[4] Xianghui Jiang, Xuanyu Liu, Yutong Wu, and Dehuai Yang. White wine quality prediction and analysis with machine learning techniques. 2023.

[5] Ji Hyung Lee, Shi Zhentao, and Zhan Gao. On lasso for predictive regression. 2021.