Clustering and Fitting

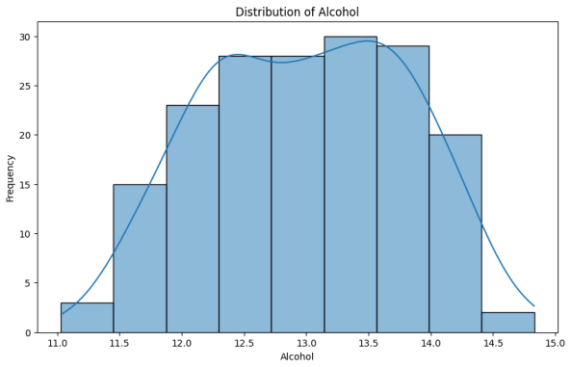
Analysis of Wines Dataset

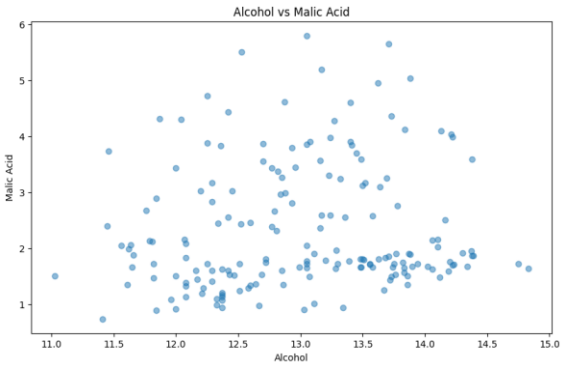
Introduction:

In this report, we perform a comprehensive analysis of a wine dataset sourced from Kaggle, which includes various chemical properties of wines. The wine dataset consists of 13 features: Alcohol, Malic_Acid, Ash, Ash_Alcanity, Magnesium, Total_Phenols, Flavanoids, Nonflavanoid_Phenols, Proanthocyanins, Color_Intensity, Hue, OD280, Proline

Histogram of Alcohol Content

The histogram below shows the distribution of the `Alcohol` content in the wine dataset. The distribution appears to be slightly right-skewed, indicating a higher frequency of wines with moderate alcohol content, centered around 13%. Most wines in the dataset have an alcohol content between 12.5% and 14%, with fewer wines having very low or very high alcohol content.



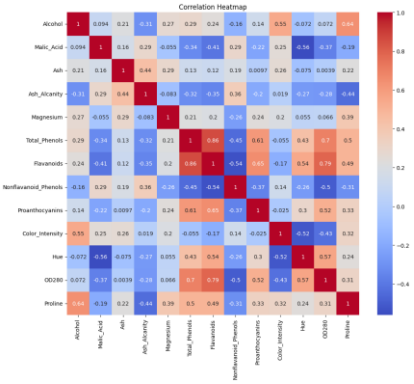


Scatter Plot of Alcohol vs Malic Acid

There is no strong linear relationship, but the scatter plot shows some clustering of data points. The data points are widely dispersed, indicating that `Malic_Acid` content varies significantly for wines with similar `Alcohol` content.

Correlation Heatmap

Strong correlations are indicated by darker colors. Notably, `Flavanoids` and `Total_Phenols` show a strong positive correlation, while `Hue` and `Color_Intensity` have a moderate negative correlation.

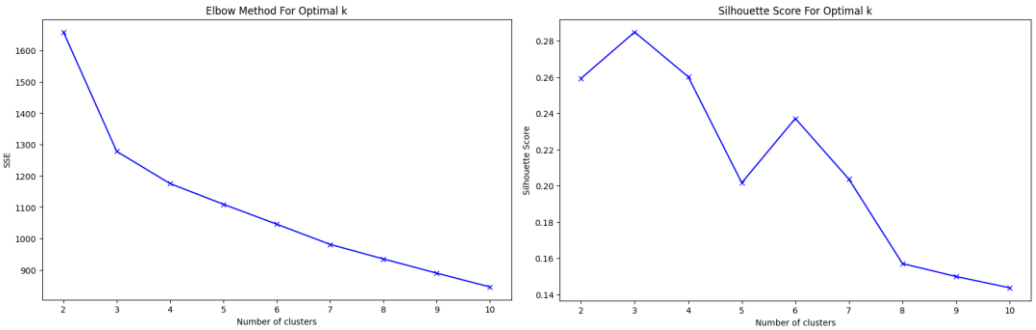


The strong correlation between `Flavanoids` and `Total_Phenols` suggests that these two features are closely related, likely because both contribute to the wine's phenolic profile. The negative correlation between `Hue` and `Color_Intensity` indicates that as the color intensity of the wine increases, its hue decreases.

Clustering Analysis

Elbow and Silhouette Method for Optimal Clusters

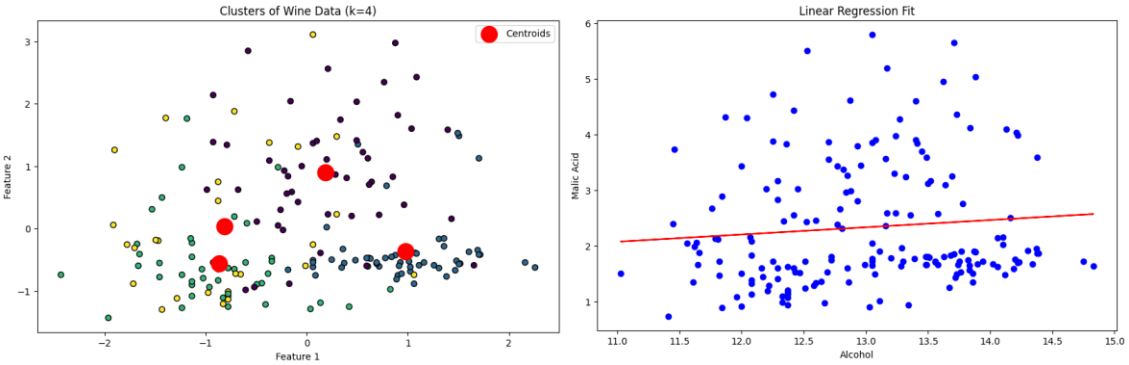
The Elbow Method and Silhouette Score plots below were used to determine the optimal number of clusters for K-Means clustering. Both methods suggest that 4 clusters are optimal for the dataset.



The elbow in the SSE plot around k=4 and the peak silhouette score at k=4 both indicate that four clusters is a suitable choice for this dataset. This suggests that the wines can be grouped into four distinct clusters based on their chemical properties.

K-Means Clustering

The clustering reveals distinct groupings in the data, indicating that the chemical properties can effectively separate wines into four groups. The centroids represent the average values for each cluster, providing a reference for the typical profile of wines within each group.



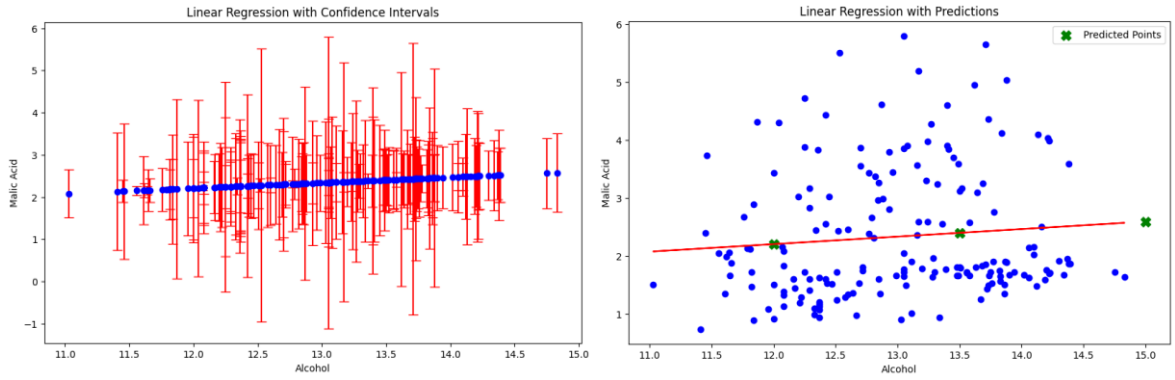
Regression Analysis

The regression line has a slight positive slope, indicating a weak positive relationship between `Alcohol` and `Malic_Acid` content. However, the data points are widely scattered, showing that `Alcohol` content alone is not a strong predictor of `Malic_Acid` content.

Predictions with Confidence Intervals

The plot below includes confidence intervals for the predictions made by the linear regression model. The red error bars represent the absolute value of the prediction errors.

The wide confidence intervals indicate high uncertainty in the predictions. This reinforces the idea that `Alcohol` content is not a strong predictor of `Malic_Acid` content, and other features should be considered for a more accurate model.



Predictions for New Data Points

Predictions were made for new data points with `Alcohol` contents of 12, 13.5, and 15. These predictions are visualized as green points on the plot below. The predicted `Malic_Acid` values for the new data points fall close to the regression line, but given the wide scatter of the original data, these predictions should be treated with caution.

Conclusion

Clustering Analysis

The K-Means algorithm identified 4 clusters as the optimal number for the given wine dataset based on the Elbow Method and Silhouette Score. The clustering visualization showed distinct groups based on the selected features, providing insights into the different types of wines.

Regression Analysis

The linear regression model showed a weak relationship between `Alcohol` and `Malic_Acid` content. The regression line and confidence intervals were plotted, and predictions for new data points were made and visualized. The wide confidence intervals highlight the high uncertainty and suggest that additional features should be considered for a more robust predictive model.

This analysis provided valuable insights into the clustering patterns of wines and the predictive relationship between `Alcohol` and `Malic_Acid` content.

Name: Yuvaraj Jammi

Student ID: 23023543

DataSet: https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering/data

Code: https://github.com/Yuva3543/AD1-clustering-and-fitting