

Name: Sai Krishna Sri Charan Boddapati

Student ID: 23034648

Github Link: github.com/CharanSSJ/Clustering-and-Fitting

1. Introduction:

This paper focuses on clustering and regression analysis on the data set consisting of 13 numerical variables related to alcohol content. Using historical plots, line plots, an elbow plot, and correlation matrices, the report determines patterns of the data, the relationship between variables, and clustering behavior. Such predictions and assessments are done using performance metrics regarding feature importance and determination of the alcohol content. The use of scatter and density plots make the results more understandable while the clustering reveals coherent relations. As depicted in this study, the research findings being employed in an array of statistical and machine learning methodologies give insights regarding relevant alcohol-related data attributes and estimate the capability for identification of high predictive value for diverse characteristics.

2. Data description:

```
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Alcohol                178 non-null    float64
1   Malic_Acid              178 non-null    float64
2   Ash                     178 non-null    float64
3   Ash_Alcanity            178 non-null    float64
4   Magnesium               178 non-null    int64
5   Total_Phenols           178 non-null    float64
6   Flavanoids              178 non-null    float64
7   Nonflavanoid_Phenols    178 non-null    float64
8   Proanthocyanins         178 non-null    float64
9   Color_Intensity         178 non-null    float64
10  Hue                     178 non-null    float64
11  OD280                   178 non-null    float64
12  Proline                 178 non-null    int64
dtypes: float64(11), int64(2)
memory usage: 18.2 KB
```

Figure 1: Data information

```
count    Alcohol    Malic_Acid    Ash    Ash_Alcanity    Magnesium \
mean    13.000618    2.336348    2.366517    19.494944    99.741573
std      0.811827    1.117146    0.274344    3.339564    14.282484
min     11.030000    0.740000    1.360000    10.600000    70.000000
25%     12.362500    1.602500    2.210000    17.200000    88.000000
50%     13.050000    1.865000    2.360000    19.500000    98.000000
75%     13.677500    3.082500    2.557500    21.500000    107.000000
max     14.830000    5.800000    3.230000    30.000000    162.000000

count    Total_Phenols    Flavanoids    Nonflavanoid_Phenols    Proanthocyanins \
mean      2.295112    2.029270    0.361854    1.590899
std       0.625851    0.998859    0.124453    0.572359
min       0.980000    0.340000    0.130000    0.410000
25%       1.742500    1.205000    0.270000    1.250000
50%       2.355000    2.135000    0.340000    1.555000
75%       2.800000    2.875000    0.437500    1.950000
max       3.880000    5.080000    0.660000    3.580000

count    Color_Intensity    Hue    OD280    Proline
mean      5.058090    0.957449    2.611685    746.893258
std       2.318286    0.228572    0.709990    314.907474
min       1.280000    0.480000    1.270000    278.000000
25%       3.220000    0.782500    1.937500    500.500000
50%       4.690000    0.965000    2.780000    673.500000
75%       6.200000    1.120000    3.170000    985.000000
max      13.000000    1.710000    4.000000    1680.000000
```

Figure 2: Summary statistic

Figure 1: Dataset attributes which shows that this database contains 13 columns and 178 instances with no missing values. It is composed mostly of data of type float64 with 2 int64 data columns. This summary is used to identify which techniques to use when analyzing the data with respect to the types and their sizes.

In Figure 2, descriptive statistics of each of the 13 numerical features of the dataset in terms of count, mean, standard deviation, percentiles and maximum values is provided

3. Visualization:

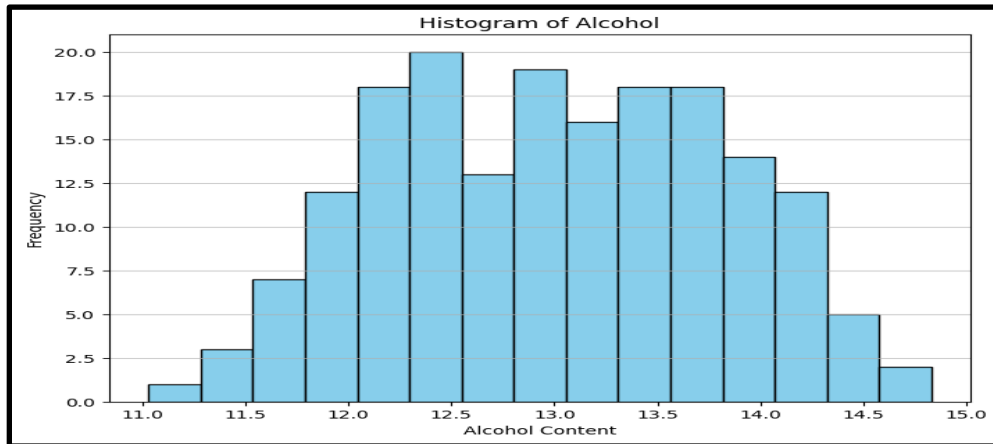


Figure 3: Histogram Of Alcohol

What is more, the histogram of alcohol content presented in figure 3 looks like a normal curve with one peak in the middle of the scale around 12.5. As for the majority of data points, they point to concentration situated between 11.5 and 14.5 alcohol content. A few are hardly located at points of 11 and many are slightly located at points of 14.5 to illustrate that the dataset has some higher and lower alcohol percentage.

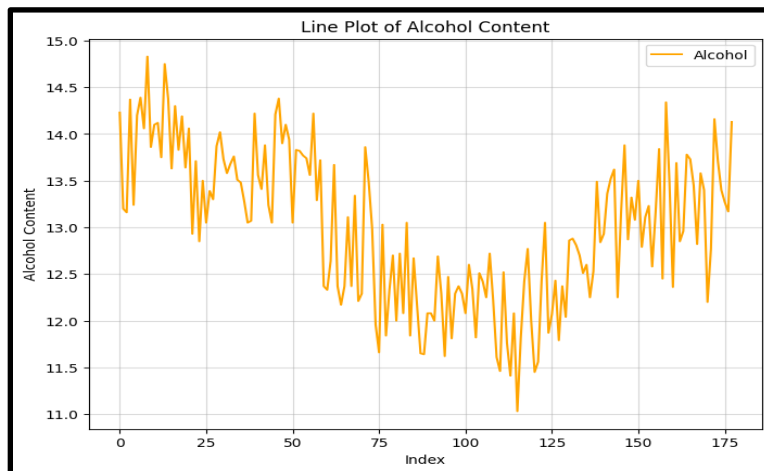


Figure 4: Line Plot of Alcohol Content

A line plot of alcohol content is shown in Figure 4 to minimize trends or patterns while showing that there are oscillations around the average value. The plot focuses on variation found in the data and helps determine if there are outliers or if the data contains anomalies. This kind of chart is useful in presenting the distribution of the data and in identifying deviant conferences in alcohol content.

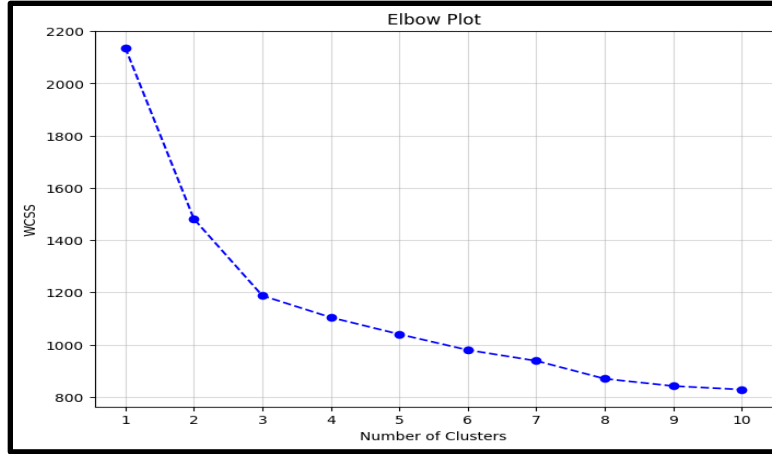


Figure 5: Elbow Plot

The elbow plot used in the current study, indicating the number of clusters appropriate for the K-means analysis is shown in figure 5. The number of clusters is on the x-axis and the sum of squares within clusters on the y-axis stands for ‘WCSS’. WCSS is an inverse function of the number of clusters and ‘elbow’ point chosen at 3 clusters represents a high balance between clustering efficiency and model simplicity.

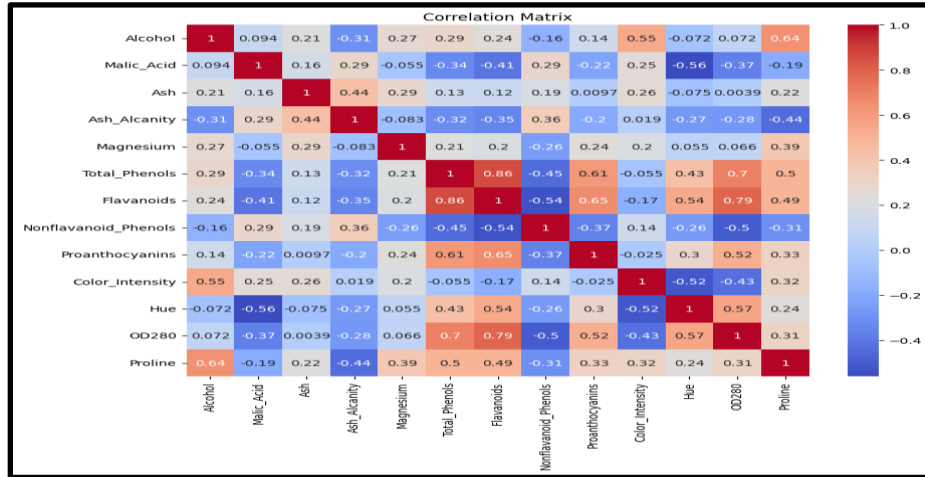


Figure 6: Correlation Matrix

Finally, Figure 6 illustrates the correlation matrix with regard to numerical features, which demonstrated pairwise proportions correlation. It is both balanced and centred along the diagonal wherein diagonal values are set at one due to auto-correlation. Colours identify correlation intensity and sign, while red means positiveness, and blue means negativeness. These are presented as strong positive co-efficient of Flavanoids and Total Phenols and strong negative co-efficient of Nonflavanoid Phenols and Proanthocyanidins, which helps in feature selection.

3.1 Discussion:

Overall, Figures 3, 4, 5, and 6 summarize the main aspects of the dataset considered in the present paper. From the histogram in figure 3, it can be deduced the central tendency of alcohol content and variability, it is observed that it followed a bell-shaped distribution. The line plot of sample-by-sample variability given in Figure 4 is also useful to identify whether there are any obvious trends in the variability. As for the elbow plot in Figure 5, the most appropriate K-means clustering is decided to be three since it is the actual number for the formation of clusters. In the feature space correlation matrix shown in figure 6, it is easier to discover strong correlations between features such as Flavonoids with Total Phenols to help in feature engineering and for guidance in clustering.

6. Conclusion:

It is found that this analysis is able to provide relevant insights by using both clustering and regression analysis to the alcohol content data set. For the relationships between features, the correlation matrix and scatter plot could be used and for the structure of clustering, the elbow plot could be conveniently used. Regression results pay more attention to feature importance and qualitatively confirm the model accuracy using measures such as R-squared and Mean Squared Error. The analysis of the predicted and actual alcohol content also proves the credibility of the models as slight disparities remain sources of improvement. The findings of this report illustrate how concepts used in analytical tools may be applied to a real-world set of data, to facilitate analytical decision making and predictive modeling in similar contexts.