THE AMERICAN
UNIVERSITY IN CAIRO

MACT 4233: Applied Multivariate Analysis - Dr. Ali Hadi

Project 2 and 3

The Application of Discriminant Analysis and Clustering Techniques on the Wine Dataset

Katia Gabriel - 900202272

Farida Simaika - 900201753

Table of Contents

Introduction

The wine industry contains various types of wines and relies solely on the characteristics that when changed produces a variety of different wine types. Each drink is coupled with a combination of characteristics that defines its nature and helps in identifying its type. The wine industry produces a wide range of different wine types, under the very big umbrella of white wines and red wines. The aim of this project is to be able to classify these wines into their types (white or red wine) based on a combination of features that best describe each group. Three methods of discriminant analysis will be discussed and implemented thoroughly throughout the course of this paper to achieve this goal. Moreover, clustering analysis will be performed to explore how the data can be grouped and whether different wine characteristics in the same group can define a new cluster of observations. Two methods of clustering analysis will be applied on the wine dataset as part of this paper's objective. Our aim is to help winemakers maximize their knowledge about the various wine types and predict whether a wine instance belongs to a certain type of wines based solely on its features.

Data Description

The data set encompassed 600 observations and 13 variables

Variable	Meaning	Unit	Type
Fixed Acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)	g / dm^3	Quantitative
Volatile acidity	the amount of acetic acid in wine, which at too high of	g / dm^3	Quantitative

	levels can lead to an unpleasant, vinegar taste		
Citric acid	found in small quantities, citric acid can add ‘freshness’ and flavor to wines	g / dm^3	Quantitative
Residual sugar	the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet	g / dm^3	Quantitative
Chlorides	The amount of salt in the wine	g / dm^3	Quantitative
Free sulfur dioxide	The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine	mg / dm^3	Quantitative
Total sulfur dioxide	Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste	mg / dm^3	Quantitative

	of wine		
Density	The density of water is close to that of water depending on the percent alcohol and sugar content	g / cm^3	Quantitative
PH	Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale	N/A	Quantitative
Sulfates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant	g / dm^3	Quantitative
Alcohol	The percent alcohol content of the wine	% by volume	Quantitative
Quality	Score between 0 and 10	N/A	Quantitative
Style	Type of wine (red or white)	N/A	Categorical

Source of the data:
https://www.kaggle.com/datasets/numberswithkartik/red-white-wine-dataset

Discriminant Analysis

Discriminant Analysis is a crucial multivariate technique that deals with the classification of each given observation into one and only one of the known groups found in the dataset at hand. This is very useful when trying to determine whether a new unseen instance is to be classified in one of the respective categories in the dataset based on which important decisions could be made in regards to this instance. There are three methods of discriminant analysis, which are: The Fisher Linear Discriminant Analysis (FLDA), The Projection Method and The Classification based on the Multinomial Distribution Method. In the course of this paper all three methods will be applied on the wine dataset in an attempt of finding the best discriminant rule that properly classifies the groups and has the smallest classification error. Moreover, all three methods will be compared to find the best performing method.

The Fisher Linear Discriminant Analysis (FLDA)

The wine dataset captures various important features such as the quality, sweetness, shelf life and acidity of previously produced wines. The dataset has a label variable which indicates the type of wine, i.e. whether each observation belongs to the white wine family (Group 1 (G1)) or the red wine family (Group 2 (G2)). After understanding the nature of the dataset, the FLDA method will be used to develop a sensible classification rule to classify a future observation into one of the two groups. However, the FLDA method assumes that G1 and G2 both follow a multivariate normal distribution and that they both have equal covariance matrices. Accordingly, both assumptions must be validated before the further implementation of the FLDA method.

Verifying the FLDA Assumptions

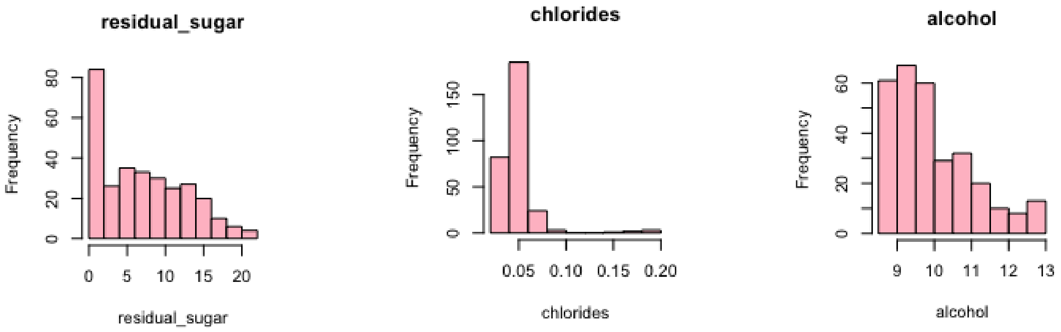
1. Checking the Multivariate Normal Distribution of the Two Groups Assumption:

After splitting the dataset into two dataset for each group, and plotting the variables in each subset, it was found that some of the variables do not follow a multivariate normal distribution. Thus, the appropriate transformations (such as log transformation or the square root) were applied. As we can see below, after transformation, the variables approximately follow a multivariate normal distribution.

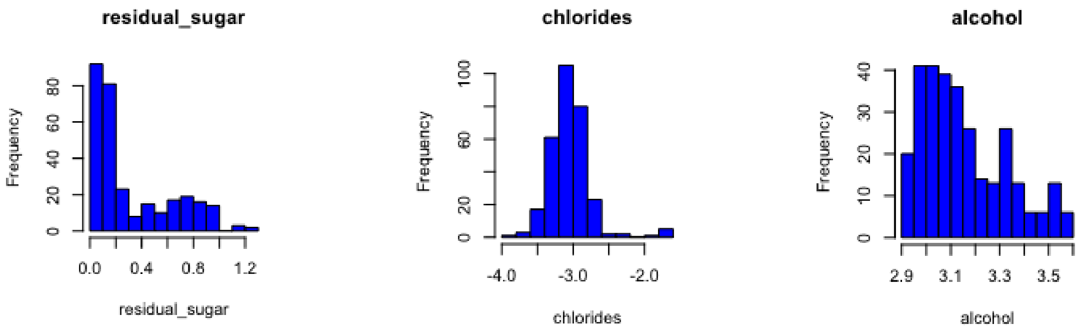
<u>For Group 1 - The White Wine Type:</u>

The variables that needed transformations were residual sugar, chlorides, density and alcohol, while the rest of the variables have proven to follow a multivariate normal distribution.

The variables residual sugar, chlorides, and alcohol before transformation:



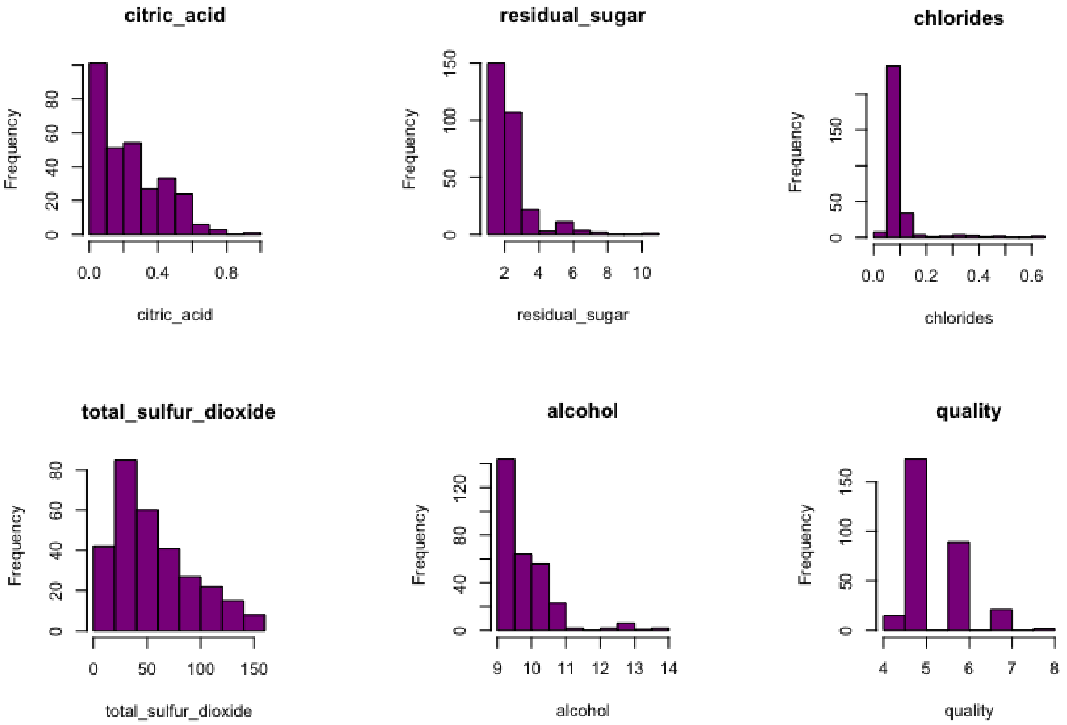
The variables residual sugar, chlorides, and alcohol after transformation:



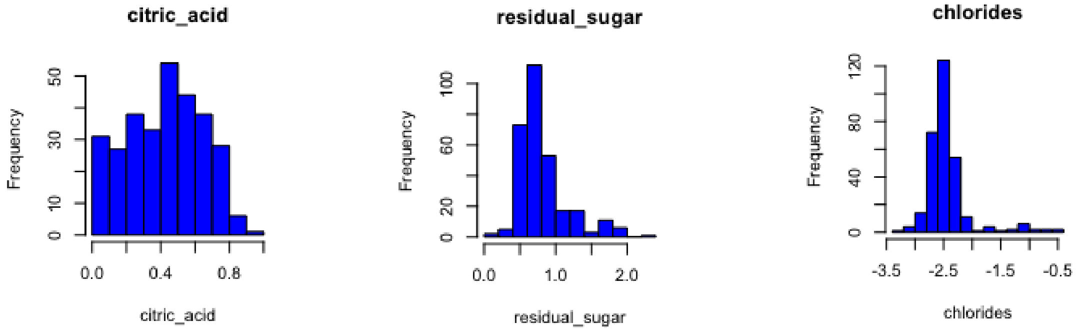
For Group 2 - The Red Wine Type:

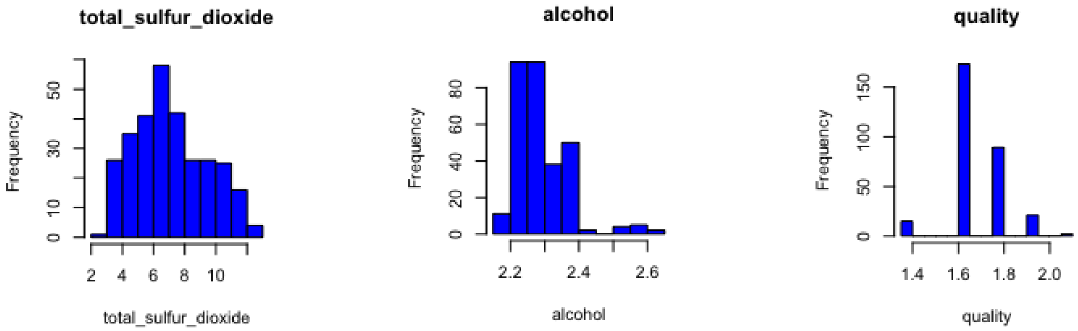
The variables that needed transformations were citric acid, residual sugar, chlorides, alcohol, total sulfur dioxide, and quality, while the rest of the variables have proven to follow a multivariate normal distribution.

The variables citric acid, residual sugar, chlorides, alcohol, total sulfur dioxide, and quality before transformation:



The variables citric acid, residual sugar, chlorides, alcohol, total sulfur dioxide, and quality after transformation:





The first assumption has been validated and both groups follow a multivariate normal distribution.

2. Checking Constant Variance Assumption:

The second assumption to be validated is that both groups in the dataset have the same variance. Accordingly, it is assumed that the covariance matrices of both groups are equal. After a thorough validation of the assumptions the FLDA will be implemented.

FLDA Implementation

Now that the assumptions have been validated, the FLDA method was implemented and yielded the following results:

```
> rslt = flda(features,class)

Fisher Linear Discriminant:


                    class   red  white
                      red   300      0
                    white     1    299
                 Error Rate = 0.1666667%
```

It is evident from the above confusion matrix that only one observation was misclassified and all other observations were classified correctly. To be able to assess the performance of the classification rule both internal and external validation will be applied.

Internal Validation

The internal validation was performed by feeding the developed discriminant rule with the same training data that was used to derive it. The following counts of correct and incorrect misclassifications were observed:

```
 class   red  white
   red   300      0
 white     1    299
Error Rate = 0.1666667
```

It can be seen from the confusion matrix that only one observation was misclassified and that the error rate is around 0.167%. However, since the rule is derived from the same data, the error might underestimate the true misclassification error.

External Validation

The external validation was performed by feeding the discriminant rule with new unseen testing data, to see how well the model behaves when encountered with future observations. The leave-one-out method was used as part of the external validation. The discriminant rule is derived using the training data with one observation left out, and is tested on the observation that has been left out. This process is repeated 599 times and the misclassification error is computed afterwards. The external validation yielded the following results:

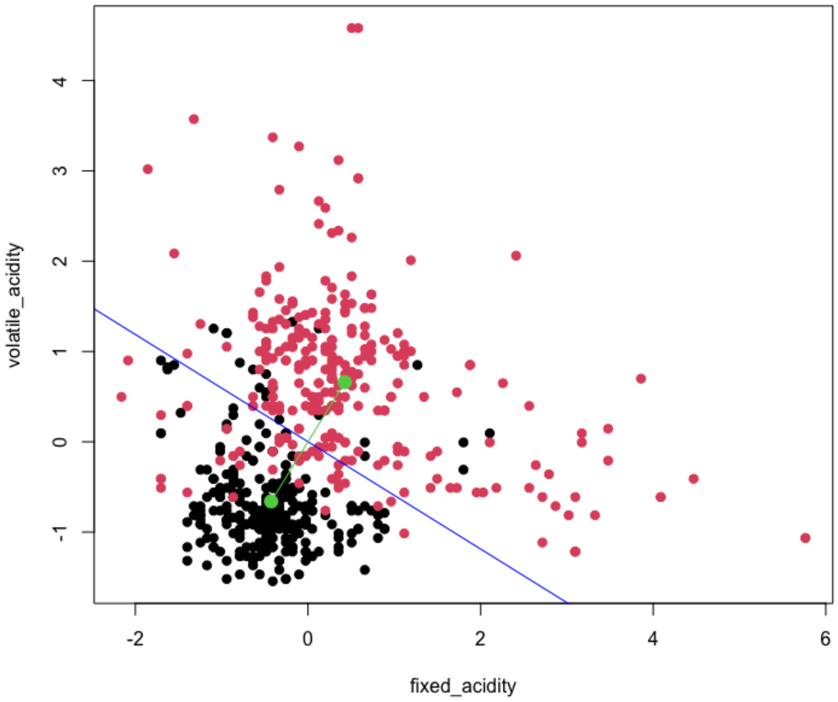
```
 class   red  white

   red   300      0
 white     1    299
```

```
Error Rate = 0.1666667
```

Both the internal and the external validations yielded the same result, which makes the developed discriminant rule a good fit for the data and has high performance and a very small misclassification error.

Since the data is multivariate, the visualization of the separability of both groups based on the 12 feature variables at the same time is hard. Instead, two variables that, when plotted together best separate the data, were chosen from the dataset and FLDA-2 was implemented and yielded the following results:



From the graph above, it is evident that both groups are separated by the blue line. Although the blue line separates the data, it does not act as a good projection because it results in a lot of overlapping points between both groups. Since the green line passes through the midpoint of the means of both groups and is almost orthogonal to the blue line, it is a good vector for projections and achieves a better separation between the two groups. It is therefore a good projection for the data. However, since the blue and the green line are both not orthogonal to each other, it indicates that both groups do not have the same variance as assumed above. Moreover, the following table shows the performance of the FLDA-2 method based on the two chosen variables from the dataset.

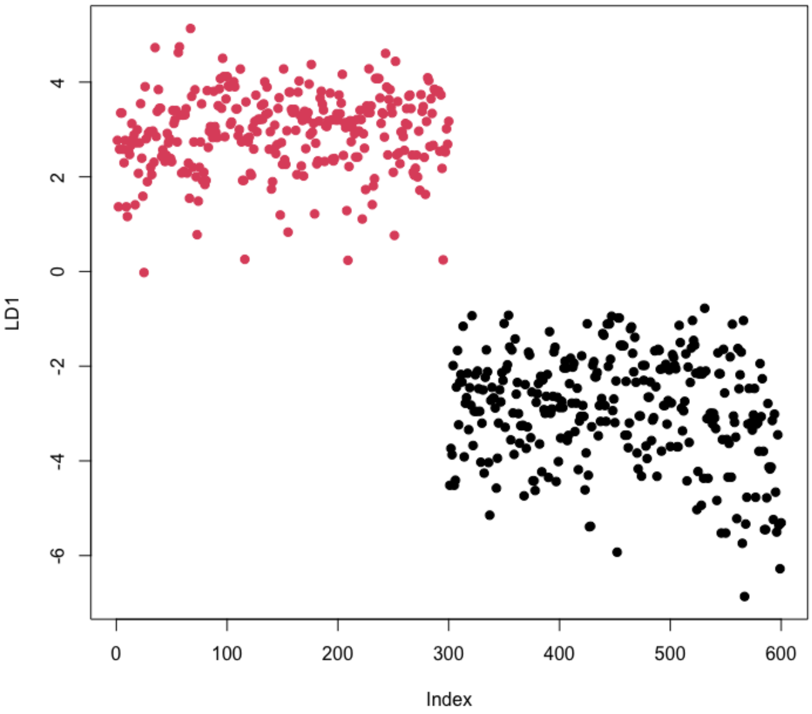
Class	Correct Classification	Incorrect Classification	Total
1	279	21	300
2	258	42	300
Total:	537	63	600

Error Rate = 10.5 %

The confusion matrix shows that 21 observations were misclassified in class 1 and 42 observations were misclassified in class 2, yielding an error rate of 10.5%. This was quite obvious from the graph since there were still few overlapping points between both groups. To find the best projection vector that best separates the two groups in the multivariate space, the projection method will be discussed.

The Projection Method

Another approach of discriminant analysis is the projection method. This method relies on the objective of finding a line or vector such that when drawn maximizes the distance between the two different groups and minimizes the distance between each observation in each group. In other words, the projection method searches for a vector such that when the sample points are projected onto it, it would yield the maximum separability. The first projection vector LD1 that corresponds to the largest eigenvalue is the best projection and is used to separate the data as depicted in the graph below:



From the graph above, it is obvious that the vector LD1 succeeded in separating the two groups completely and has therefore achieved maximum separation. The vector has allocated all members of group 1 to positive values and all members of group 2 to negative values.

The Classification based on Multinomial Distribution Method

The multinomial distribution method is a special case of the Generalized Linear Models which can be used as a discriminant analysis tool. Accordingly, after applying this method to the wine data set the following results were obtained:

```
     results
class   red  white
  red   299      1
white     2    298
  Error Rate = 0.5%
```

According to the above result the classification based on the multinomial distribution method yielded a 0.5% error rate, as 3 observations were misclassified. The discriminant rule resulting from this method seems to fit the nature of the data. However, to be able to assess the performance of the classification rule both internal and external validation will be applied.

Internal Validation

The internal validation was performed by feeding the developed discriminant rule with the same training data that was used to derive it. The following counts of correct and incorrect misclassifications were observed:

```
     results
class   red  white
  red   299      1
white     2    298
 Error Rate = 0.5%
```

It can be seen from the confusion matrix that 3 of the observations were misclassified and that the error rate is 0.5%. However, since the rule is derived from the same data, the error might underestimate the true misclassification error.

External Validation

The external validation was performed by feeding the discriminant rule with new unseen testing data, to see how well the model behaves when encountered with future observations. The leave-one-out method was used as part of the external validation. The discriminant rule is derived using the training data with one observation left out, and is tested on the observation that has been left out. This process is repeated 599 times and the misclassification error is computed afterwards. The external validation yielded the following results:

```
           results
      class   red  white
        red   297      3
      white     3    297
        Error Rate = 1%
```

Looking at the results of the external validation, it is obvious that more observations were misclassified. A total of 6 observations were misclassified, yielding an error rate of 1%. This underscores the fact that the internal validation underestimates the true misclassification error. Since the external validation tests the discriminant rule based on new unseen instances, it is more accurate in evaluating the rule and hence it could be concluded that the classification based on multinomial distribution method has an overall error rate of 1%

Conclusion

After performing both the FLDA method and the classification based on the multinomial distribution method, it is evident that the FLDA method best suits the nature of the wine dataset. By comparing the internal validation of the both methods it is evident that the FLDA was able to classify all observations correctly in their respective groups and only failed in misclassifying one instance, yielding a very low error rate of 0.1667%. However, that of the multinomial distribution method, misclassified 3 instances and yielded an error rate of 0.5% which is higher than the error rate of the FLDA method. Since the internal validation usually underestimates the error rate of misclassification, it is crucial to compare the external validation of both aforementioned methods. Looking at the FLDA, it is evident that both the internal and external validation yielded the same results, which makes the developed discriminant rule a good fit for the data and has high performance and a very small misclassification error. However, the external validation of the multinomial method yielded an even higher error of 1% due to the misclassification of 6 instances. After comparing the performance of both methods it becomes clear that the Fisher Linear Discriminant Analysis is most suitable for the wine dataset and achieves good classification results.

Cluster Analysis

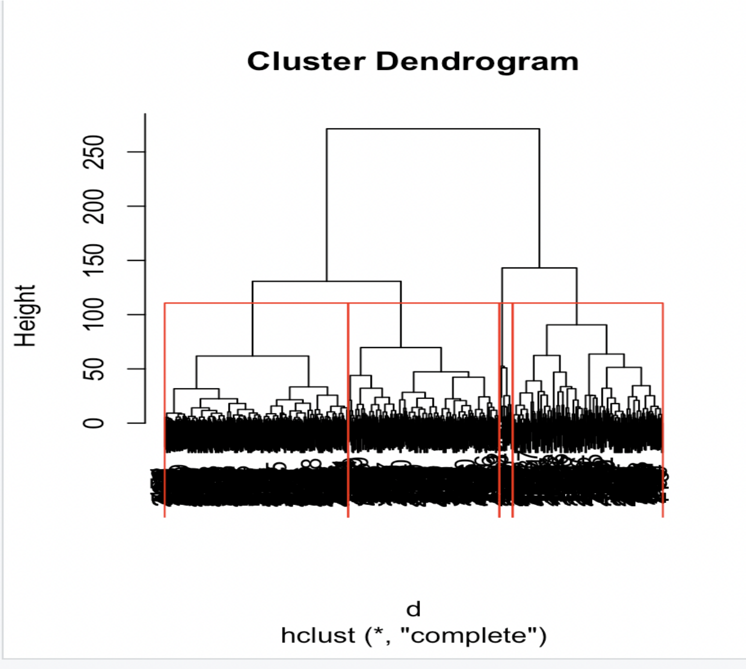
Clustering is a data analysis technique used to group similar objects or data points into clusters based on their distance or similarity. Two popular clustering algorithms are K-means and Hierarchical clustering. Both algorithms aim to minimize inter-cluster distances and maximize intra-cluster distances. Minimizing inter-cluster distances means that data points in different clusters are as different as possible, while maximizing intra-cluster distances means that data points within the same cluster are as similar as possible. This helps to ensure that clusters are meaningful and useful for data analysis. Cluster analysis is an unsupervised learning technique , the number of clusters in the data is unknown.

Hierarchical Clustering

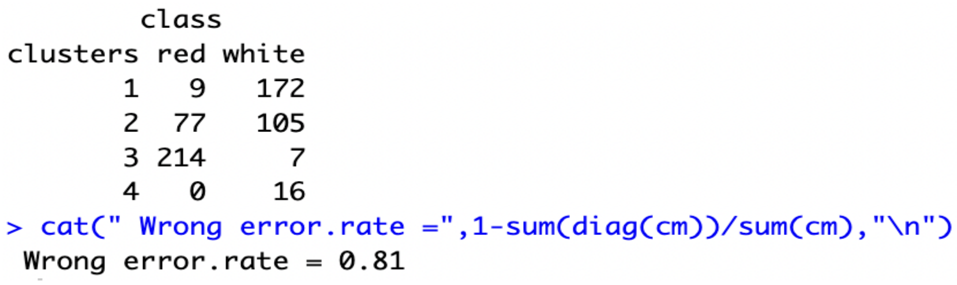
Hierarchical clustering is a type of unsupervised learning that does not require the knowledge of any predefined number of clusters in the data. The hierarchical clustering algorithm builds a tree-like structure of clusters, which is also known as a dendrogram. The algorithm works by starting with each data point as its own cluster. Then, it iteratively merges the two closest clusters until all the data points belong to a single cluster at the base of the dendrogram. The distance between different observations is calculated using a distance metric, such as the Euclidean or Manhattan distances. Both distances measure the dissimilarity between the attributes of the data points.

One of the critical components of the algorithm is the distance metric and linkage method used to calculate the distance between the data points and clusters respectively. In this project, hierarchical clustering using different distance metrics (Manhattan and Euclidean distances) and different linkage techniques (complete and single linkage) will be implemented on this dataset.

Hierarchical Clustering: Euclidean Distance and Complete Linkage



The above dendrogram was obtained by using the Euclidean distance between observations and the complete linkage between clusters. The dendrogram shows that there are subjectively four clusters in this dataset at a threshold of 100. The algorithm grouped together similar data points by iteratively merging the two closest clusters until all data points belong to a single cluster at the base of the dendrogram. It can be seen that the fourth cluster contains relatively few observations. Careful analysis of the data shows that most of the data points classified as outliers had significantly high fixed acidity which increased their overall quality grades.



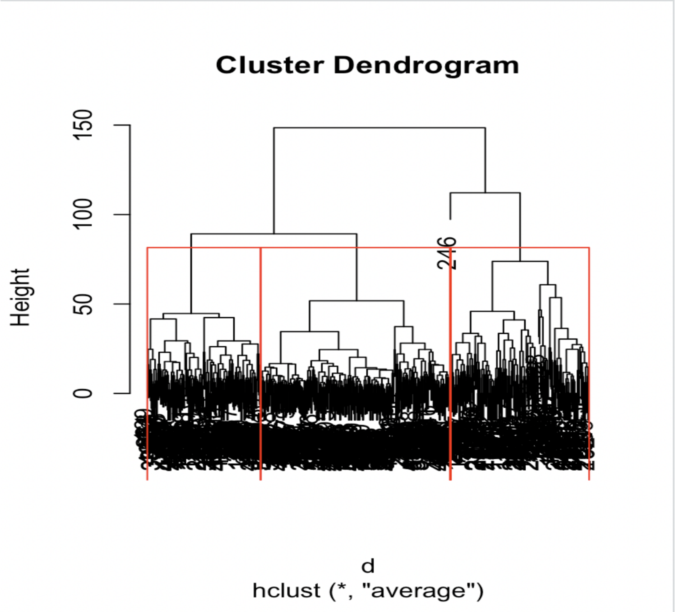
```
         class
clusters red white
       1   9   172
       2  77   105
       3 214     7
       4   0    16
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.81
```

This output from R shows the number of observations that belong to each cluster (1,2,3 and 4) and each class (red and white wines). Cluster 1 has a total of 9 observations that belong to the red wine class and 172 observations that belong to the white class. Cluster 2 has a total of 77 observations that belong to the red class and 105 observations that belong to the white class. Cluster 3 has 214 observations that belong to the red class and 7 observations belonging to the white class. Clusters 1,2 and 4 appear to be dominated by the white wine class, while cluster 3 appears to be dominated by the red wine class. The error rate is quite high at 0.81, indicating that the model is not performing well on the given dataset. The specific cause of the high error rate may be due to a poor choice of features.

Further analysis of the data showed that the 16 observations in the fourth clusters are in fact outliers. These points had low sulfate quantities which degraded the overall quality of these wines. Sulfates are important additives in wine because they prevent its oxidation and are essential to preserving the wine.

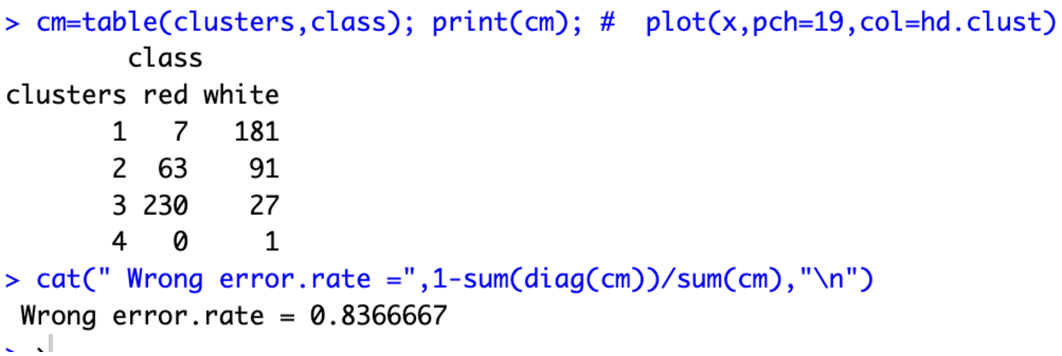
An R^2 value of 88% usually shows that the hierarchical clustering implemented on the data is somewhat satisfactory. It suggests that the data points are closely clustered together and the model is doing a good job of explaining the variance in the data. However, in this case, due to the significant error rate, a high value of R^2 is misleading. A dataset with a few outliers that significantly deviate from the rest of the data (which is the case in our dataset) or a larger number of clusters can lead to an inflated R^2 value.

Hierarchical Clustering: Manhattan Distance and Average Linkage



›

Based on the above dendrogram obtained using the Manhattan distance between instances and average linkage between clusters, it appears that there are four clusters in the dataset at a threshold of 60. The fourth cluster contains only one observation (observation 246) which indicates that this point might be an outlier. Further analysis of the data led to the conclusion that observation 246 is in fact an outlier. The observation presented low sulfate quantities which degraded the overall quality of these wines.

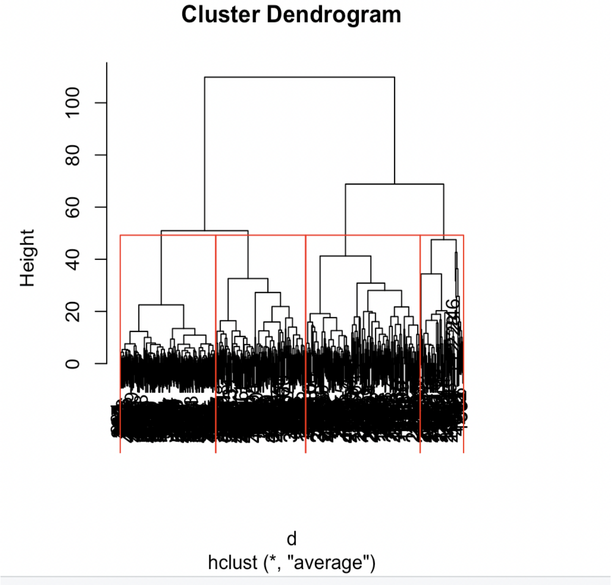


```
> cm=table(clusters,class); print(cm); #  plot(x,pch=19,col=hd.clust)
        class
clusters red white
       1   7   181
       2  63    91
       3 230    27
       4   0     1
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.8366667
```

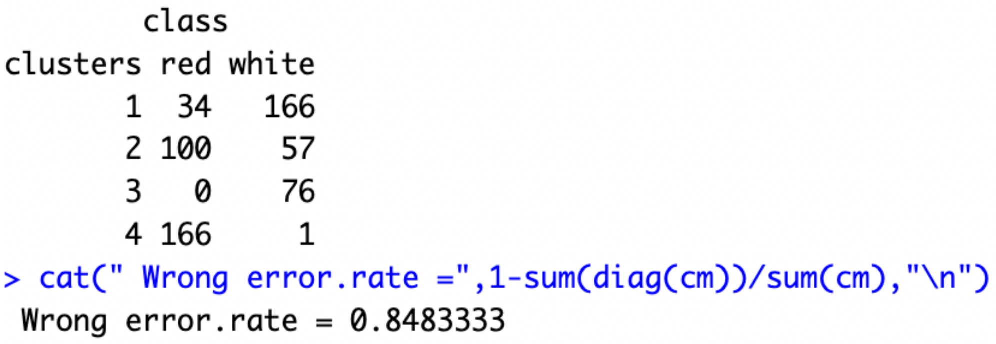
This R output displays the number of observations in each cluster (1, 2,3 ands 4) and their class (red and white wines). Within Cluster 1, there are 7 observations in the red wine class and 181 observations in the white wine class, while Cluster 2 has 63 observations in the red wine class and 91 observations in the white wine class. Cluster 3 has 230 observations in the red wine class and 27 observations in the white wine class while cluster 4 contains a unique observation in the white wine group. The confusion matrix suggests an error rate of around 84% which indicates that the model is not performing well on the dataset. The root cause of this high error rate could be due to an improper selection of features or distance measure.

An R^2 value of 84% usually shows that the hierarchical clustering implemented on the data is somewhat satisfactory. It suggests that the data points are closely clustered together and the model is doing a good job of explaining the variance in the data. However, in this case, due to the significant error rate, a high value of R^2 is misleading. A dataset with a few outliers that significantly deviate from the rest of the data (which is the case in our dataset) or an increased number of clusters can lead to an inflated R^2 value.

Hierarchical Clustering: Euclidean Distance and Average Linkage



The above dendrogram was obtained by using the Euclidean distance between observations and the complete linkage technique between clusters. The dendrogram shows that there are subjectively four clusters in this dataset at a threshold of 50. As can be seen above, the fourth cluster contains a few number of observations. These observations need to be investigated in order to assess whether or not they are outliers. As previously mentioned these observations had significantly high fixed acidity which increased their overall quality grades . The high acidity levels are one of the reasons behind the presence of outliers in the red wine group.



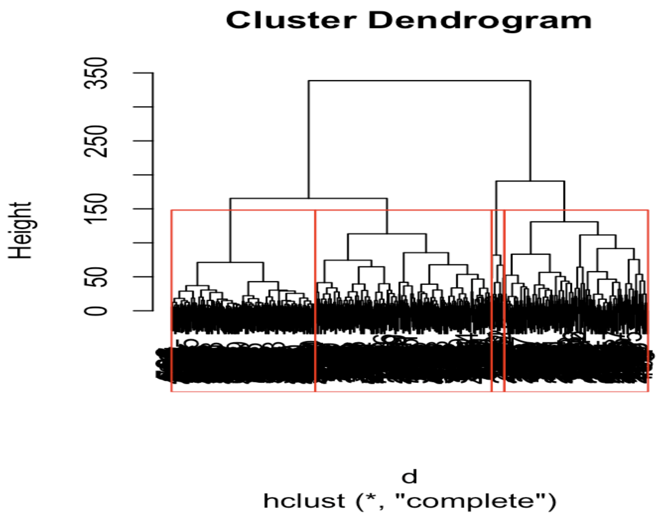
```
        class
clusters red white
       1  34   166
       2 100    57
       3   0    76
       4 166     1
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.8483333
```

This R output displays the number of observations in each cluster (1, 2,3 ands 4) and their class (red and white wines). Within Cluster 1, there are 34 observations in the red wine class and 166 observations in the white wine class, while Cluster 2 has 100 observations in the red wine class and 57 observations in the white wine class. Cluster 3 has 76 observations in the red wine class while cluster 4 contains a unique observation (observation 246) in the white wine group and 166 observations in the red wine group. As previously mentioned, observation 246 is a wine with significantly low sulfates quantities and therefore a bad quality wine.

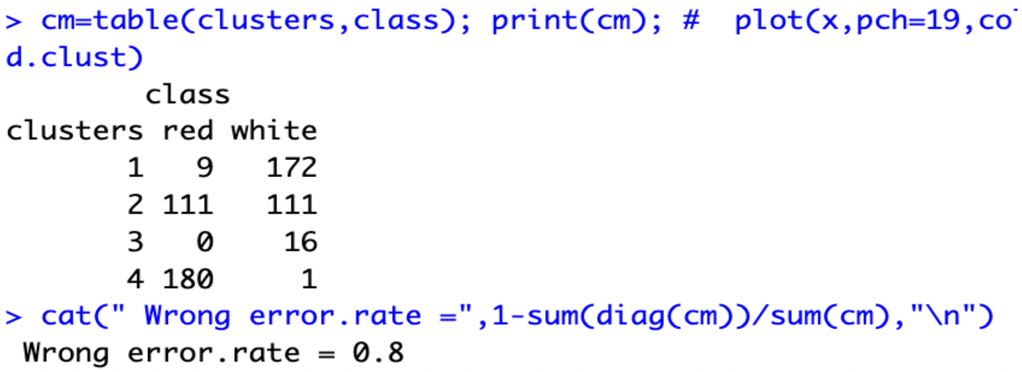
The confusion matrix suggests an error rate of around 84% which indicates that the model is not performing well on the dataset. The root cause of this high error rate could be due to an improper selection of features or distance measure.

An R^2 value of 89% usually shows that the hierarchical clustering implemented on the data is somewhat satisfactory. It suggests that the data points are closely clustered together and the model is doing a good job of explaining the variance in the data. However, in this case, due to the significant error rate, a high value of R^2 is misleading. A dataset with a few outliers that significantly deviate from the rest of the data (which is the case in our dataset) or an increased number of clusters can lead to an inflated R^2 value. A high R^2value may also suggest overfitting, meaning the model is too complex and may not generalize well to new data.

Hierarchical Clustering with Manhattan Distance and Complete Linkage



The above dendrogram was obtained by using the Manhattan distance between observations and complete linkage between clusters. The dendrogram shows that there are subjectively four clusters in this dataset at a threshold of 150. As can be seen above, the fourth cluster contains a few number of observations.



```
> cm=table(clusters,class); print(cm); #  plot(x,pch=19,co
d.clust)
        class
clusters red white
       1   9   172
       2 111   111
       3   0    16
       4 180     1
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.8
```

The confusion matrix suggests an error rate of around 80% which indicates that the model is not performing well on the dataset. However, this is the lowest error rate reached.

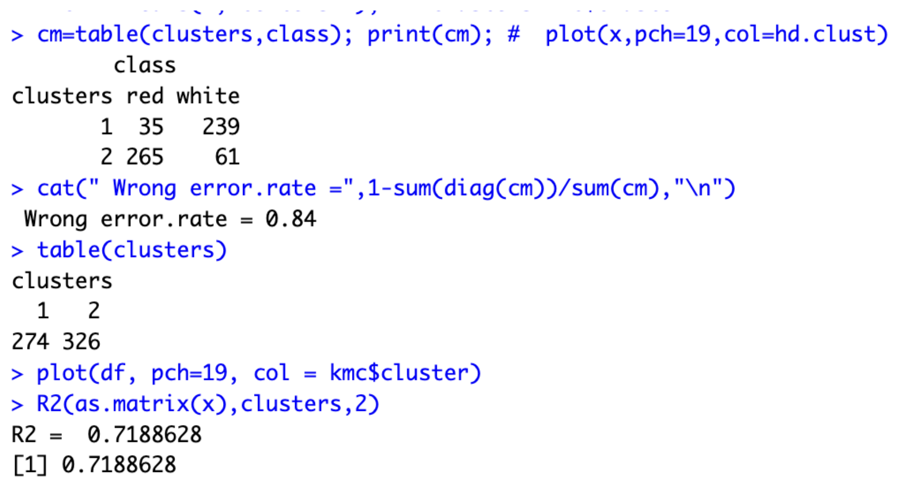
This R output displays the number of observations in each cluster (1, 2,3 ands 4) and their class (red and white wines). Within Cluster 1, there are 9 observations in the red wine class and 172 observations in the white wine class, while Cluster 2 has 111 observations in the red wine class and 111 observations in the white wine class. Cluster 3 has 76 observations in the white wine class while cluster 4 contains a unique observation (observation 246) in the white wine group and 180 observations in the red wine group. As previously mentioned, observation 246 is a wine with significantly low sulfates quantities and therefore a bad quality wine.

An R^2 value of 87% usually shows that the hierarchical clustering implemented on the data is somewhat satisfactory. It suggests that the data points are closely clustered together and the model is doing a good job of explaining the variance in the data. However, in this case, due to the significant error rate, a high value of R^2 is misleading. A dataset with a few outliers that significantly deviate from the rest of the data (which is the case in our dataset) or an increased unnecessary number of clusters can lead to an inflated R^2 value.

K-Means

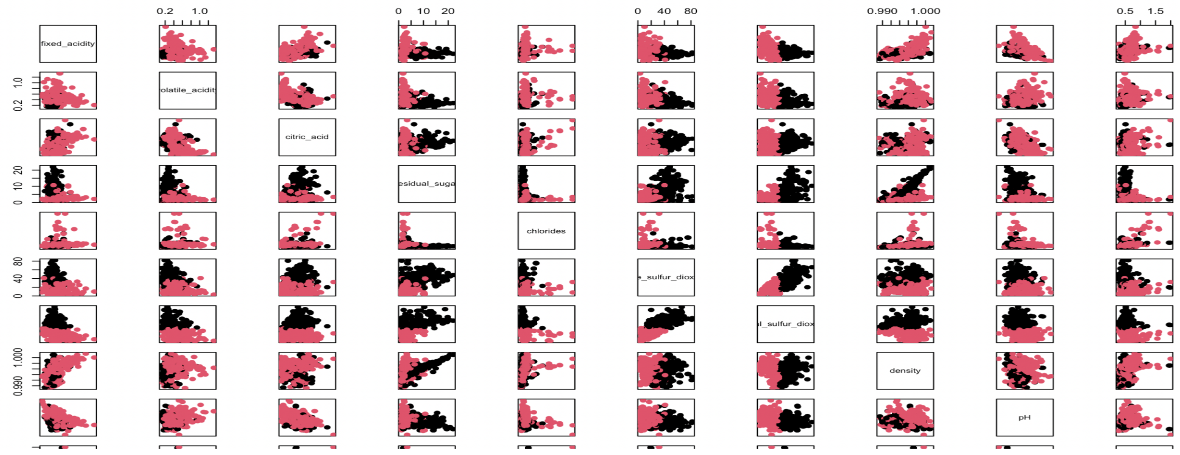
K-means clustering is a simple yet powerful technique that partitions data points into k clusters based on similarity in their features. The algorithm works by first randomly initializing k cluster centroids, and then iteratively assigning each data point to the nearest centroid and updating the centroids to minimize the sum of squared distances between data points and their assigned centroid. The process continues until convergence is reached and the centroids no longer change. One of the main advantages of k-means clustering is its scalability and efficiency. It can handle large datasets with high-dimensional feature spaces and is relatively fast compared to other clustering algorithms. The K-means clustering technique requires the knowledge of the number of clusters in advance. In this dataset, there are supposedly two clusters (red wines and white wines).

K-Means using 2-Means



```
> cm=table(clusters,class); print(cm); #  plot(x,pch=19,col=hd.clust)
        class
clusters red white
       1  35   239
       2 265    61
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.84
> table(clusters)
clusters
  1   2
274 326
> plot(df, pch=19, col = kmc$cluster)
> R2(as.matrix(x),clusters,2)
R2 =  0.7188628
[1] 0.7188628
```

The k-means clustering results for the red and white wine dataset show that there are two clusters with varying numbers of data points assigned to each. Cluster 1 includes 274 data points, cluster 2 has 326 instances. The high error rate of 84% suggests that the clusters may not be well-separated, or that there may be outliers that are affecting the clustering results.

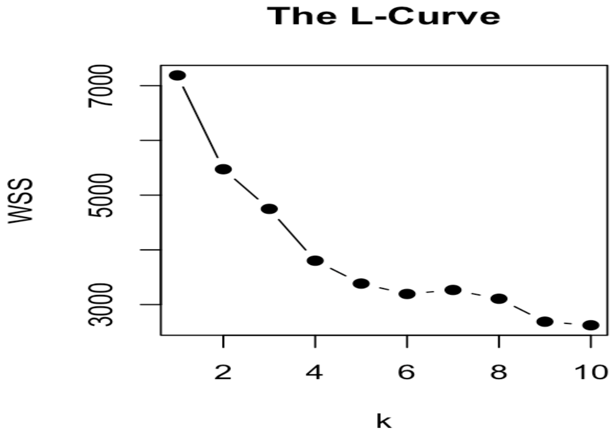


As can be seen above, the clusters are overlapping with minimal to no distinction between the two clusters.

In this case the high R^2 value is misleading, this inflated value is probably due to an increased number of clusters when implementing k-means algorithm.

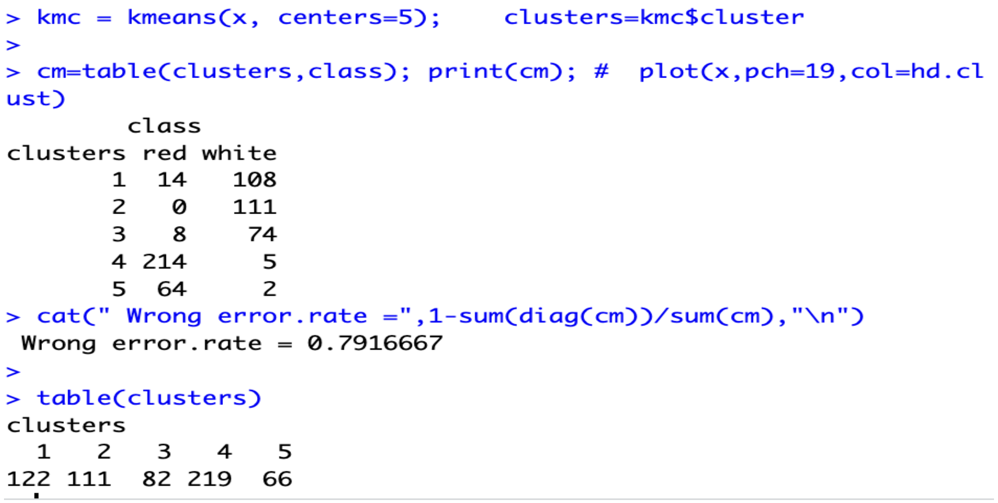
Choosing Optimal Value of K

The elbow curve is a useful tool for choosing the optimal value of k in k-means clustering. It plots the sum of squared distances between data points and their assigned centroid for different values of k. As the value of K increases, the sum of squared distances decreases, since each data point is closer to its assigned centroid. However, beyond a certain point, adding more clusters does not significantly reduce the sum of squared distances.



The elbow curve shows the "elbow point" where the decrease in sum of squared distances starts to level off, indicating the optimal value of k for the dataset. The optimal value of k seems to be at 5.

K-Means using 5-Means



```
> kmc = kmeans(x, centers=5);     clusters=kmc$cluster
>
> cm=table(clusters,class); print(cm); #  plot(x,pch=19,col=hd.cl
ust)
        class
clusters red white
       1  14   108
       2   0   111
       3   8    74
       4 214     5
       5  64     2
> cat(" Wrong error.rate =",1-sum(diag(cm))/sum(cm),"\n")
 Wrong error.rate = 0.7916667
>
> table(clusters)
clusters
  1   2   3   4   5
122 111  82 219  66
```

The k-means clustering results for the red and white wine dataset show that there are five clusters with varying numbers of data points assigned to each. Cluster 1 includes 122 data points, cluster 2 has 111 data points, cluster 3 has 82 data points, cluster 4 has 219 data points, and cluster 5 has 66 data points. The confusion matrix shows the number of data points in each cluster that belong to either red or white wine. For example, cluster 1 has 14 points that belong to red wine and 108 points that belong to white wine. On the other hand, cluster 4 has 214 points that belong to red wine and only 5 points that belong to white wine. The error rate calculated using the confusion matrix is 0.7916667, which indicates that the clustering model is not performing well. This suggests that the clusters may not be well-separated, or that there may be outliers that are affecting the clustering results.

Goodness of Fit of K-Means

The R-squared value of 0.5095122 for k-means clustering suggests that approximately 50% of the variance in the data can be explained by the clustering model. This means that the model is doing a moderately good job of identifying patterns and grouping similar data points together. However, it also indicates that there is still a significant amount of unexplained variance in the data that the clustering model has not captured.

Conclusion

The wine dataset has a very high error rate for both hierarchical clustering and K-means. This suggests that the data is too complex or noisy for these algorithms to effectively identify meaningful patterns in the data. Clustering algorithms work by identifying patterns and similarities in the data, but if the data is too complex or noisy, it can be difficult to accurately group the data points into meaningful clusters. In this case, the algorithm may end up assigning data points to the wrong cluster, resulting in a high error rate. In such cases, it may be useful to explore other clustering algorithms or preprocessing techniques to better prepare the data for clustering. It is also possible that there is no clear structure within the dataset that can be effectively captured by clustering. A poor choice of variables (from the creators of this data set) may also be the reason behind the high error rate. The features used to cluster the data may not be informative enough to accurately differentiate between the different types of wine.