葡萄酒品質分析之研究

吳杰峰
M11123015
國立雲林科技大學資訊管理所

梁淯愷
M11123036
國立雲林科技大學資訊管理所

王敬翔
M11123051
國立雲林科技大學資訊管理所

劉昱辰
M11123057
國立雲林科技大學資訊管理所

中華民國 111 年 12 月 14 日

摘要

品酒文化無論古今都是人們生活模式的樂趣之一，許多人藉由品酒獲得生活的樂趣。但酒的品質優劣程度是關乎於每個品酒人的興致，除指定資料集資料集外，自選題使用 Wine Quality Data Set，本組利用多種演算法如 K-means 演算法、DBSCAN 以及 n-clusters 等演算法，藉由資料的預測與分析判斷葡萄酒的優劣性質分析，期盼藉由科技的力量，讓葡萄酒之品質優劣能夠分析的透徹，未來更期盼運用在各種不同加工食品品質優劣之預測

關鍵字: 葡萄酒、資料分析、決策樹、K-means

Research on Wine Quality Analysis

WU, JIE-FENG
M11123015
Department of Information Management
National Yunlin University of Science and Technology

LIANG, YU-KAI
M11123036
Department of Information Management
National Yunlin University of Science and Technology

WANG, CHING-HSIANG
M11123051
Department of Information Management
National Yunlin University of Science and Technology

LIU, YU-CHEN
M11123057
Department of Information Management
National Yunlin University of Science and Technology

Abstract

Wine tasting is one of the most important pleasures of people, and many, many wine tastings are the joys and pleasures of life. However, the quality of wine is related to each wine quality data set. This group uses the algorithm algorithm, dbscan and n-clusters and other n-clusters, etc. The power of science and technology makes the excellent quality of grape wine products thoroughly analyzed. In the future, we look forward to using it in the prediction of the excellent quality of various processed foods

Key words: wine quality, data analysis, decision tree, K-means

一、動機

品酒文化無論古今都是人們生活模式的樂趣之一，許多人藉由品酒獲得生活的樂趣。但酒的品質優劣程度是關乎於每個品酒人的興致，但隨著近年來科技的進步，我們已經可以利用科技的力量來預測或分析葡萄酒品質的優劣與好壞，因此，本組將葡萄酒的品質定位為此次預測的研究主題。

二、目的

這兩個數據集與葡萄牙“Vinho Verde”葡萄酒的紅色和白色變體有關。有關詳細信息。由於隱私和物流問題，只有物理化學（輸入）和感官（輸出）變量可用（例如沒有關於葡萄類型、葡萄酒品牌、葡萄酒售價等的數據）。

這些數據集可以被視為分類或回歸任務。這些類別是有序的且不平衡（例如，普通葡萄酒比優秀或差的葡萄酒多得多）。離群值檢測算法可用於檢測少數優質或劣質葡萄酒。此外，我們不確定所有輸入變量是否相關。所以測試特徵選擇方法可能很有趣。除了自選資料集以外，還有一個指定資料集，並利用 K-means 演算法等進行酒的優劣品質的預測分析，讓品酒人以及未來想從事加工食品研究分析的研究人員有數據的佐證及繼續研究的依據。

三、方法

3.1 實作說明

在資料集一中，使用 K-means_cluster，先匯入資料集，並找到 K 值的績效，在資料集分成三群和算出資料群的密度。使用 hierarchical_clustering，先顯示 Scatter plot 再找出 Sepal Length 和 Sepal Width 的關係，並透過上述的關係找出 petal_width 和 sepal_width 的關係，並用 Pairplot 顯示每個可能的數字列之間創建一個聯合圖，若數據框很大，則需要一段時間，用 Violin Plot 可以發現 Iris setosa 的 sepal width 大於 Iris virginca，最後是 Iris versicolor。Boxplot 顯示平均 sepal length 長度大於其他特徵，其次是 petal length 和 sepal width 還有最小的 petal width，接下來找出使用 dendrogram 找出最佳的聚集數字，最後找出最佳值後套入 Agglomerative 裡，並算出 purity score。使用 DBSCAN，先用 Pairplot 在每個可能的數字列之間創建一個聯合圖，接下來再切割 data 分成三群進行類據分析，並通過輪廓分數進行聚類評估，最後算出每個 K 值的結果和 purity score。

在資料集二，使用 K-means 中先創建一個標準化的物件，找到最佳的 k 值後並帶入模組中執行。使用 hierarchical_clustering，在 n_clusters 迴圈係數找出最佳 k 值並用手肘法選出 n-clusters，在用 n-clusters 值算出執行模組所耗費的時間，並算出 Hierarchical Clustering。使用 DBSCAN，透過 DBSCAN(eps=5, min_samples=2).fit(df)創建模組，並算出執行時間。

四、實驗

4.1 前置處理

在資料集二中，先將兩個資料集合併，在將類別特徵和屬性特徵分清楚，在查詢資料集中有無空值。

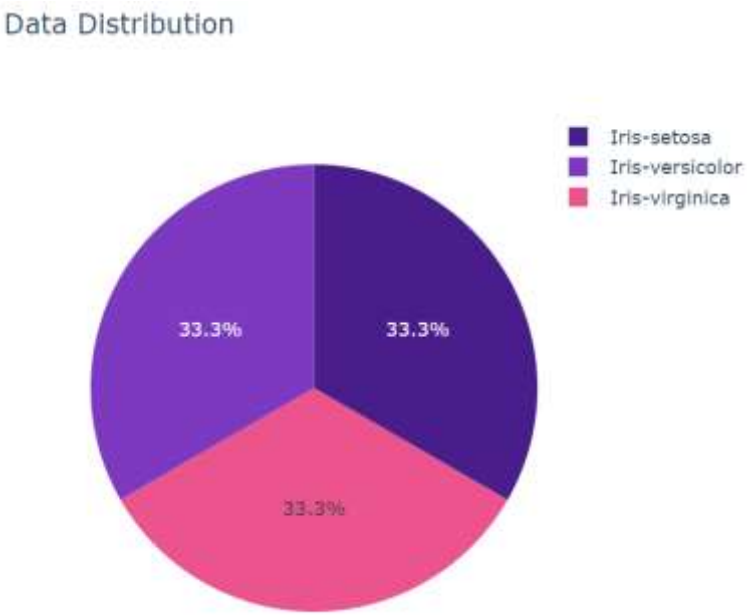
4.2 實驗設計

在資料集一，使用 K-means_cluster，透過迴圈算出 KMeans(n_clusters=i , max_iter=300)最佳的 n_clusters 值，並透過 go.Scatter 來將三個群組視覺化，最後用混淆矩陣算出資料的密集度。使用 hierarchical_clustering，透過 scatterplot()來發現對於 Iris setosa，我們可以看到 2 個變量的關係，也就是 Sepal Length and Sepal Width 之間存在正線性關係，用 Label Encoding 將標籤轉換成數字形式，將其轉換成機器可讀的形式，然後讓演算法可以更好決定如何操作這些 label，接下來把前面 scatterplot 的 X 和 Y 值套入 sch.dendrogram(sch.linkage(X, method = 'ward'))中顯示階層式分群的階層樹，而階層樹中的最佳聚集為 2，但是在左邊是沒有聚類的實際散點圖，不同聚類內有數據點衝突，右側是預測的聚類，數據點被準確分類並聚類為 2 組。但在視覺上我們可以發現可以是另一個集群的數據點，所以將 n_clusters 改為 3 在做出新的模型發現，左邊沒有聚類的實際散點圖，不同聚類內有數據點衝突。右側是預測的聚類，數據點被準確分類並聚類為 3 組，最後用 Purity 指標算出分群品質。使用 DBSCAN，先將資料集用 sns.pairplot(iris_df, hue='Species')視覺化，接下來在將資料集分成三群並進行聚類，在通過輪廓分析進行聚類評估並新增到數據中，最後再透過每個群組的 dbscan 來畫出 scatterplot，並使用 Purity 指標衡量分群品質。

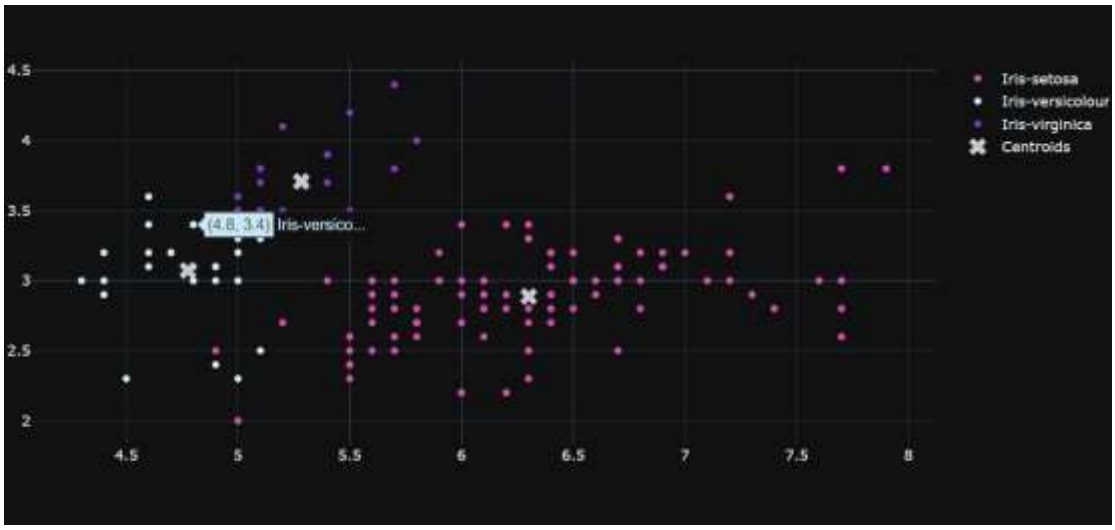
在資料集二中，使用數據中的 columns 做出 correlation matrix，並透過 MinMaxScaler()函數創建一個標準化的物件，針對 df 的資料進行計算並產生標準化資料，使用 elbow 方法輸出構建模型，也就是 n_clusters 為 2，而透過 end-start 來算出 n_clusters 為 2 和 10 的 kmeans 模組執行時間，並產生 n_clusters 為 2 個 Silhouette Score，接下來用 n_clusters 的增加來 K-Means 的績效越來越好，並找出適合數據的 n_clusters 值。為不同數量的集群創建 KMeans 實例，在使用 KMeans 實例創建 SilhouetteVisualizer 實例並實例視覺化，在顯示 n-clusters 和 WCSS 的各個 K 值的效益。接下來使用 groupby()將 k_means.labels_做分群，並顯示 alcohol 的資料、創建 alcohol 和 total sulfur dioxide 的二維圖和三維圖，最後 Aggmodel.labels_ 顯示矩陣、silhouette_score() 算出此模型的耗費時間和 metrics.v_measure_score()的績效，最後在用 dendrogram 來顯示階層式分群的階層樹。而 DBSCAN 則是將兩個樣本之間的最大距離設為 5 和將一個點視為核心點的鄰近樣本數設為 2，在算出執行此模型所耗費時間和使用 silhouette_score 算

出輪廓分析的結果，透過 etrics.v_measure_score 來衡量聚類性能的評估

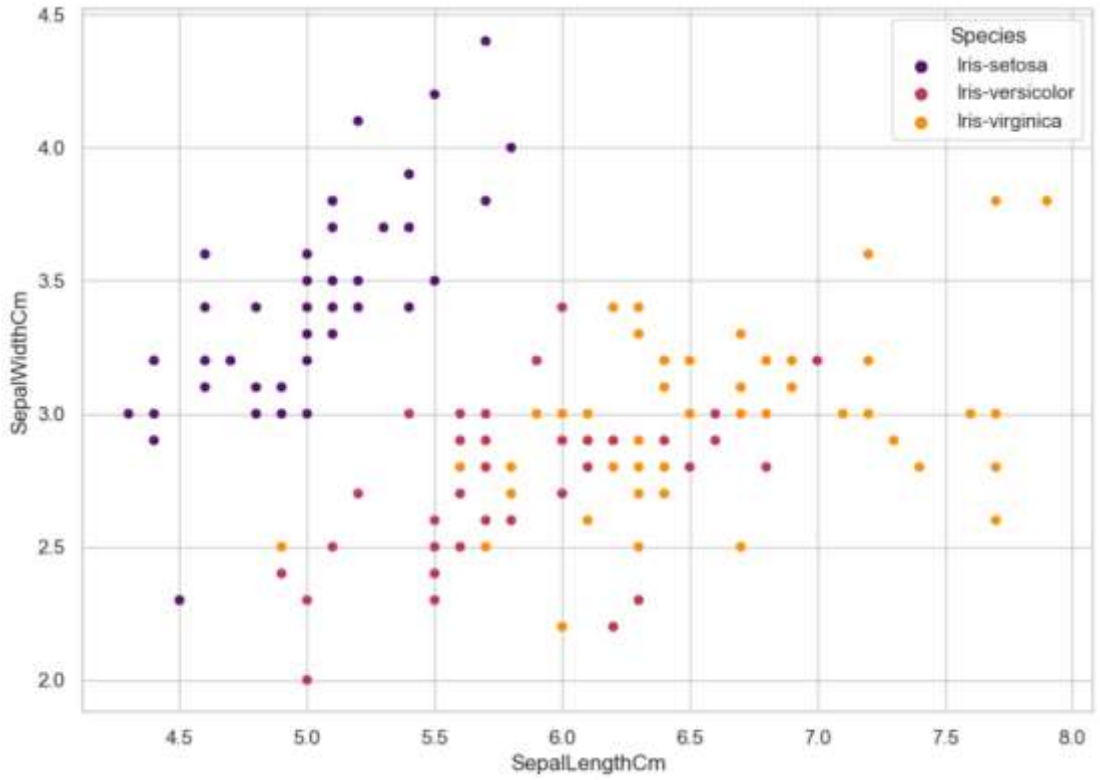
4.3 實驗結果



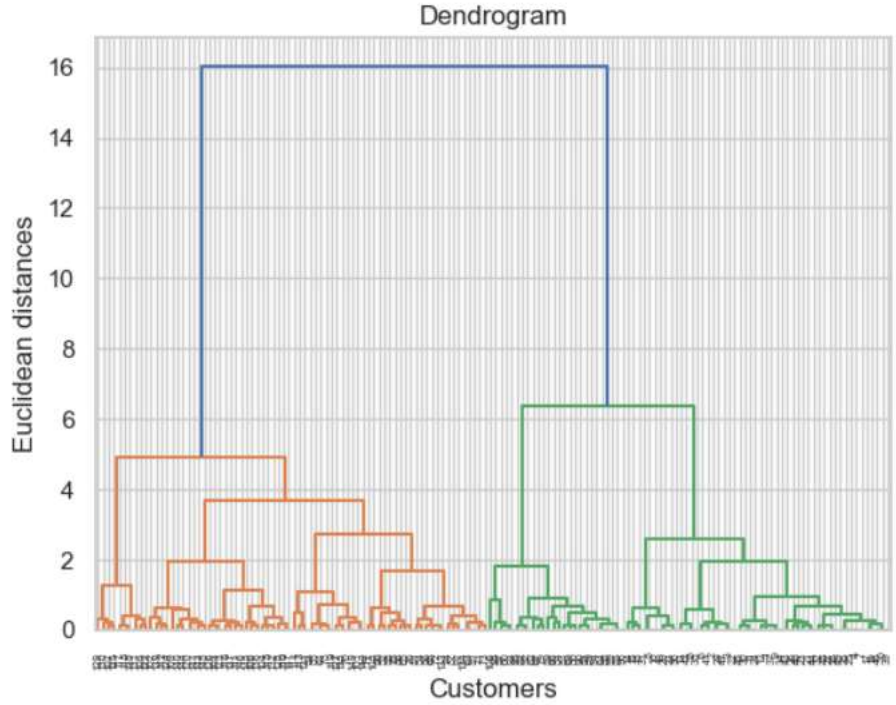
圖一、數據分佈



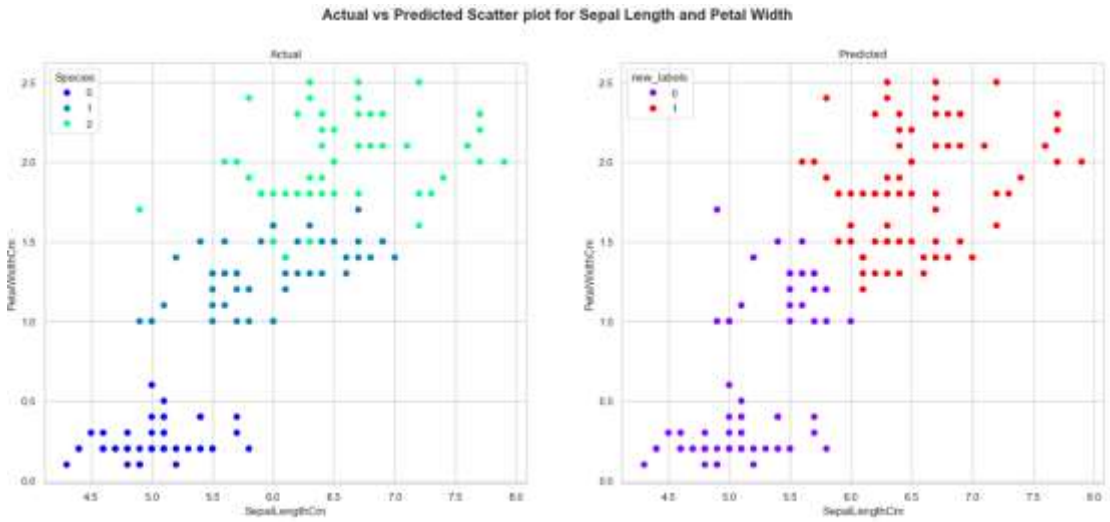
圖二、K-means 的視覺化評估結果



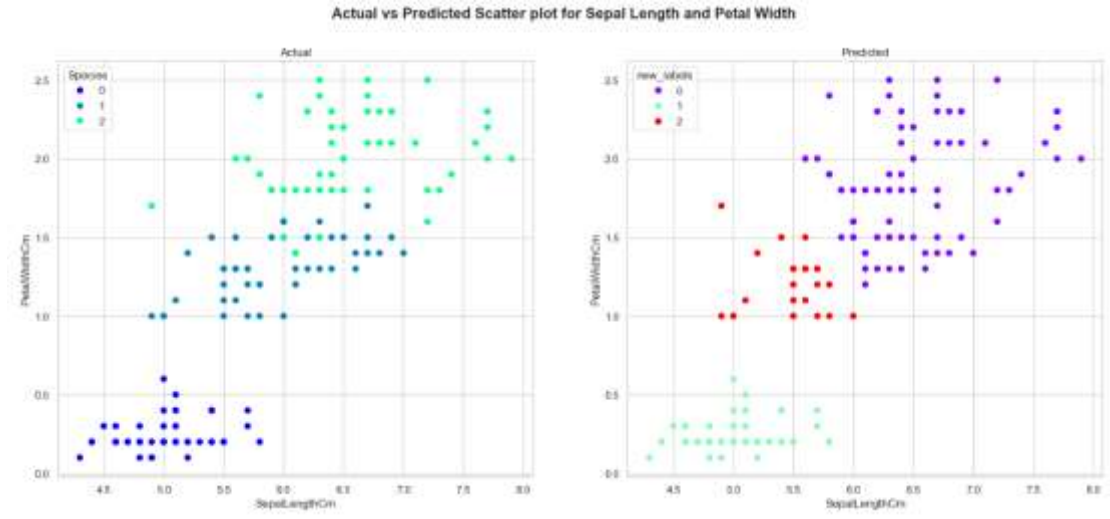
圖三、SepalLengthCm 和 SepalWidthCm 的散佈圖



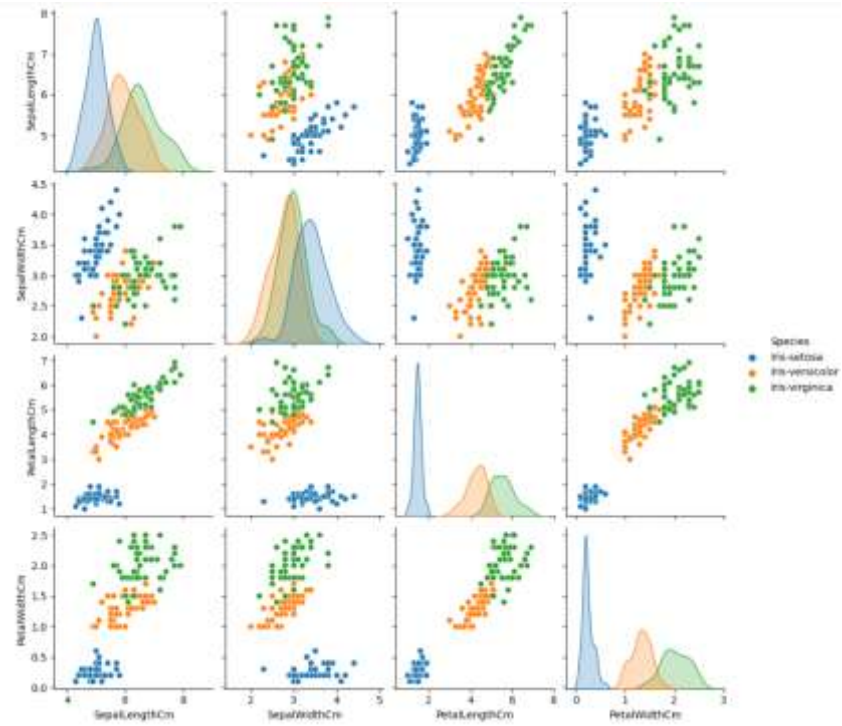
圖四、階層式分群的階層樹



圖五、當 n_clusters 為 2 時的分佈



圖六、當 n_clusters 為 3 時的分佈



圖七、特徵間的關係矩陣

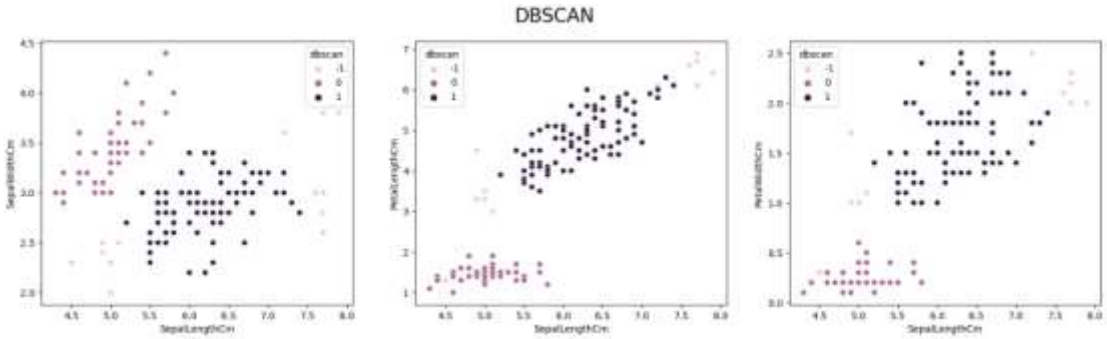
dbscan	-1	0	1
Species			
0	1	49	0
1	4	0	46
2	8	0	42

圖八、Species 和 dbscan 的交叉表

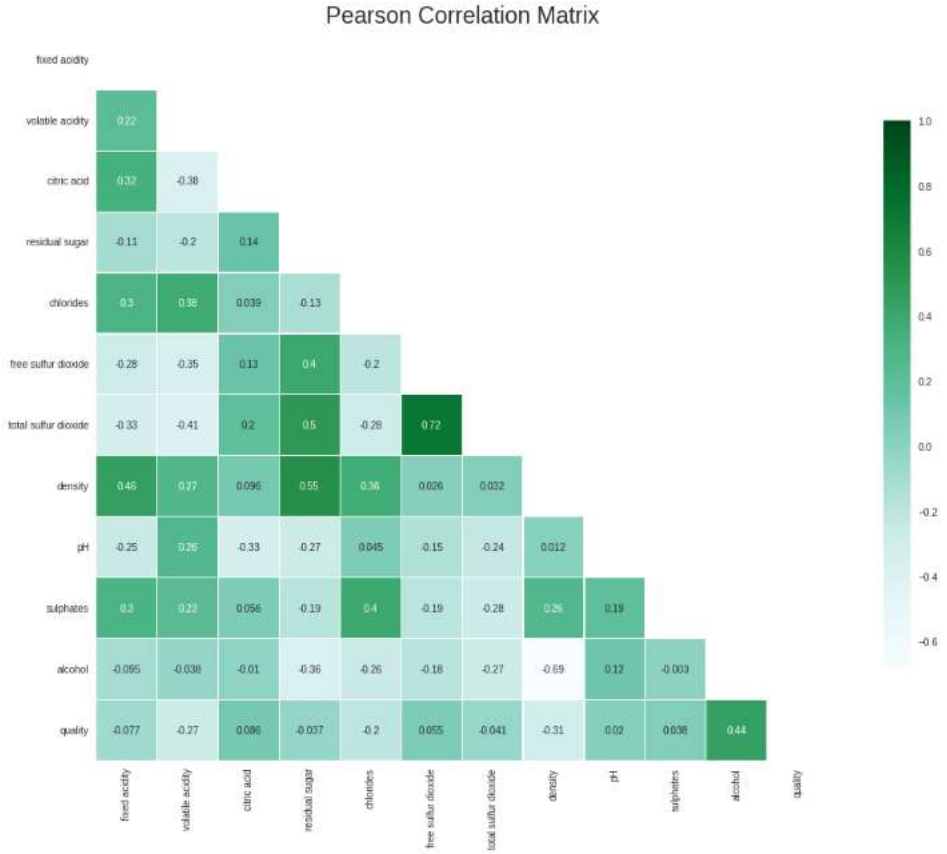
```
DBSCAN Silhouette Analysis Score : 0.5419178830436095
```

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	dbscan	silhouette_dbscan
0	5.1	3.5	1.4	0.2	0	0	0.875929
1	4.9	3.0	1.4	0.2	0	0	0.844998
2	4.7	3.2	1.3	0.2	0	0	0.854948
3	4.6	3.1	1.5	0.2	0	0	0.835301
4	5.0	3.6	1.4	0.2	0	0	0.871778

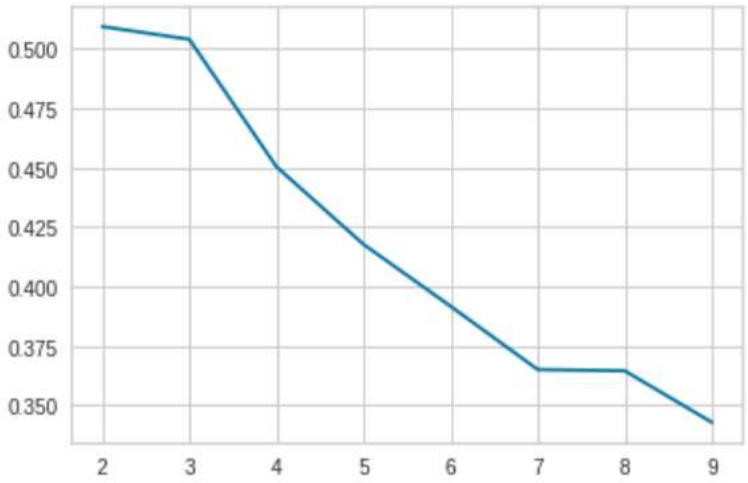
圖九、通過輪廓分數進行聚類評估的結果



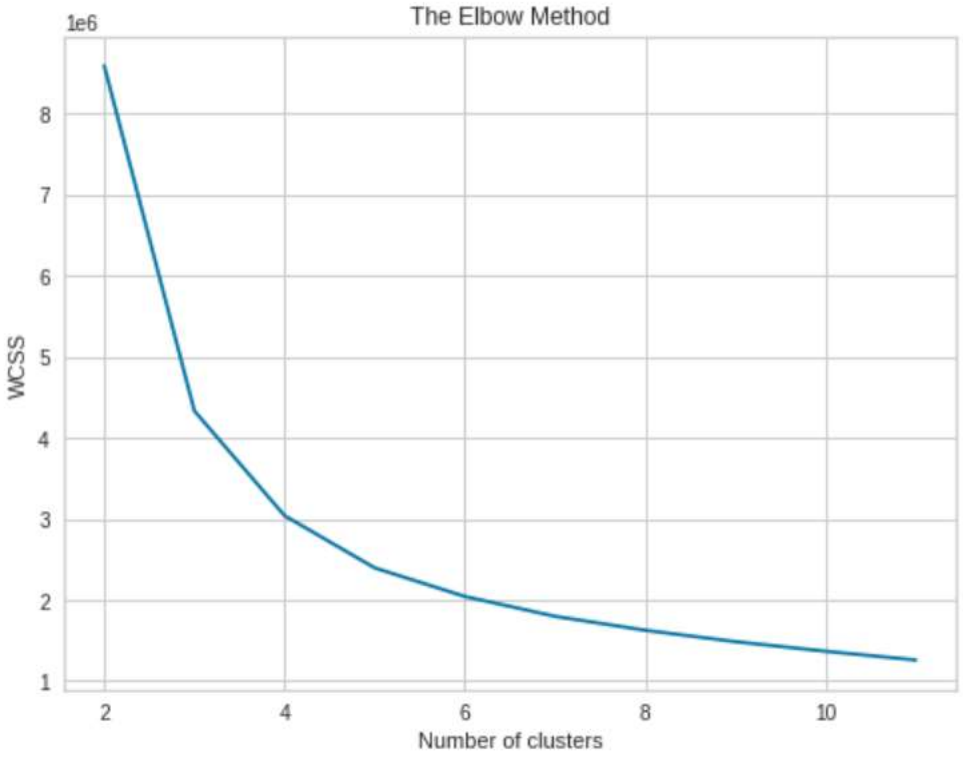
圖十、DBSCAN 各群組的分佈



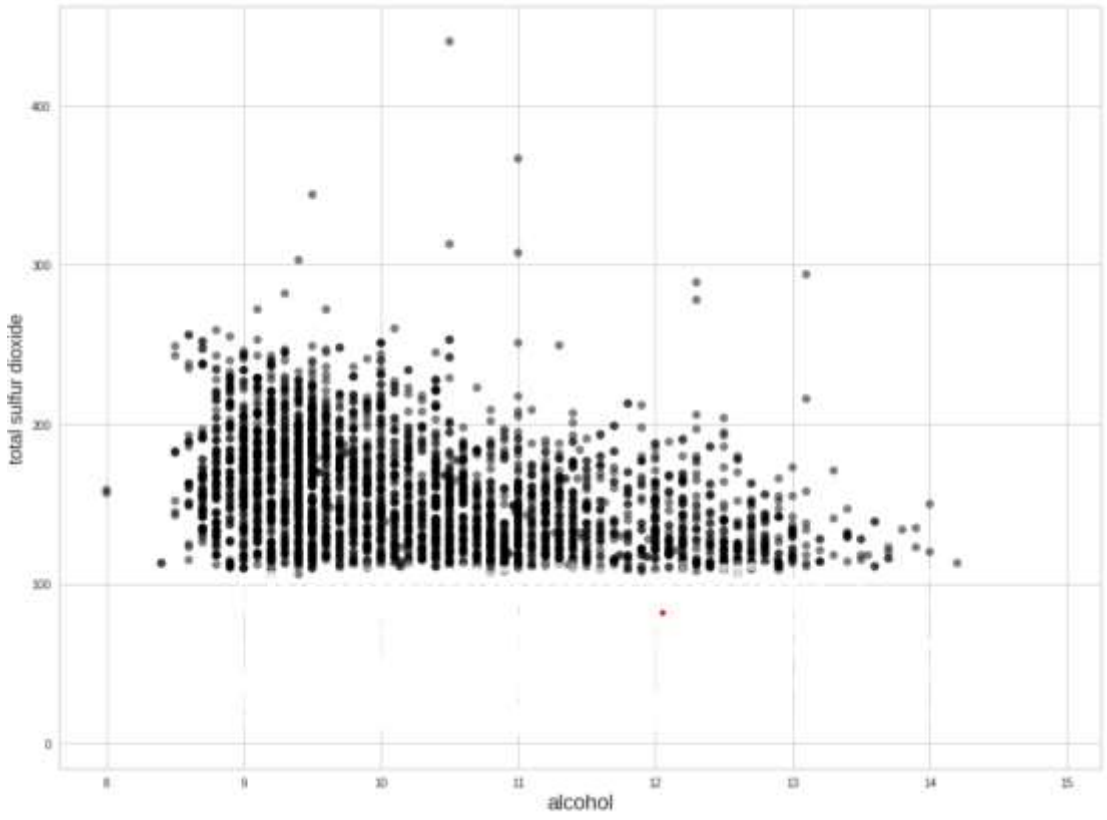
圖十一、相關係數矩陣



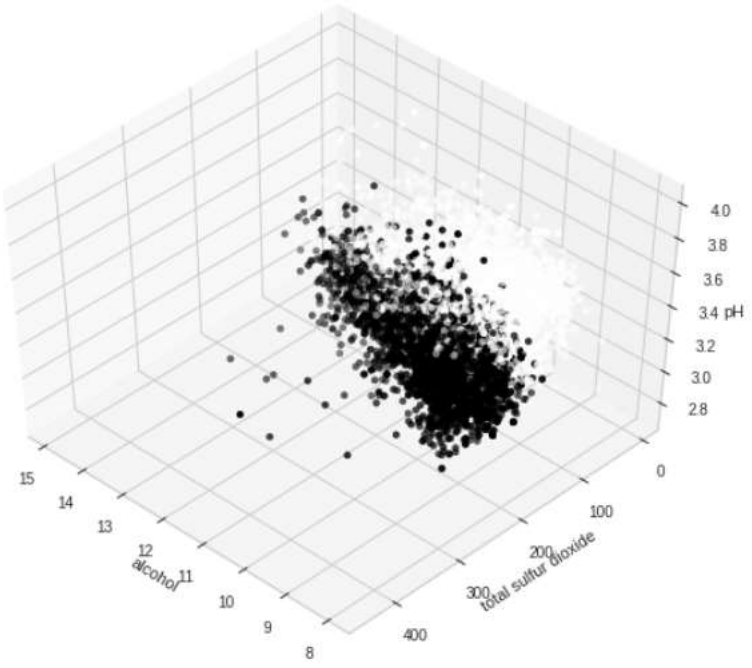
圖十二、n_clusters 迴圈係數



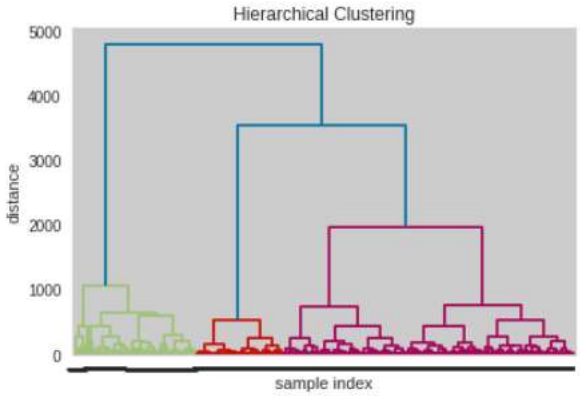
圖十三、手肘法選出 n-clusters



圖十四、alcohol 和 total sulfur dioxide 的二維圖



圖十五、alcohol 和 total sulfur dioxide 和 pH 的三維圖



圖十六、階層式分群的階層樹

五、結論

	耗費時間	Silhouetter Score
K-means	0.848	0.509
階層式分群	1.712	0.281
DBSCAN	0.121	0.281

從本組利用多種演算法的比較與訓練數據來觀察，K-means 有著較為其他二者較優質的分析品質。因此本組認為 K-means 演算法較為適合做為此次預測分析的演算法。

六、參考資料

Wine Quality Data Set; 2009-10-07

Retrieved from: https://archive.ics.uci.edu/ml/datasets/Wine+Quality