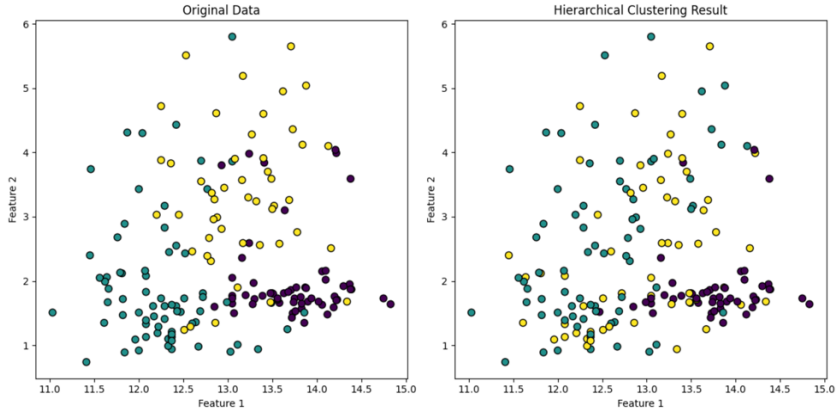
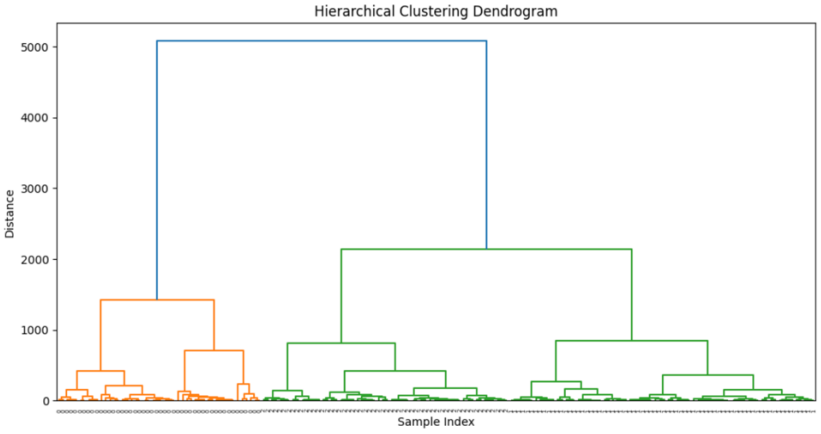
Clustering

312112009

Hierarchical Clustering

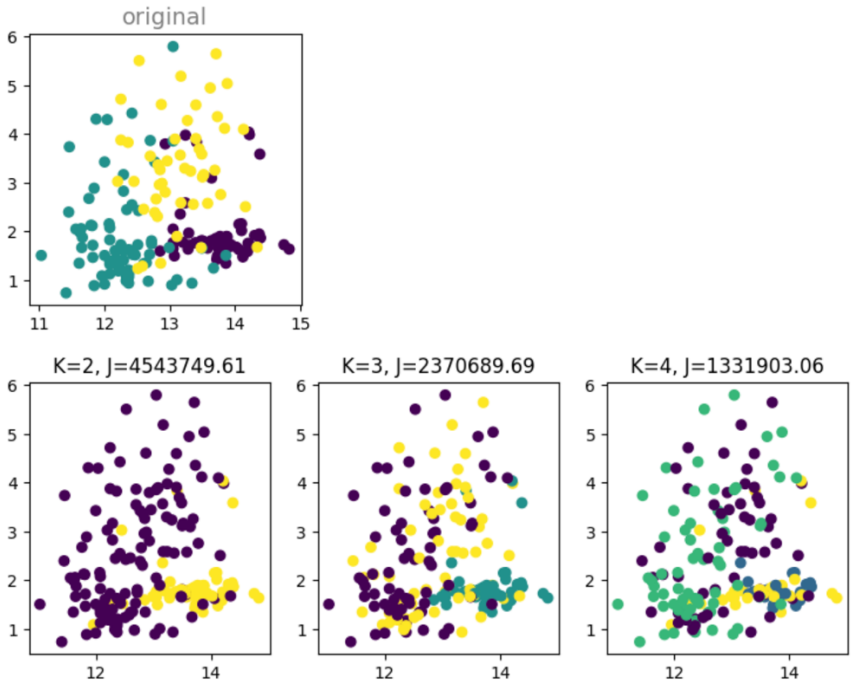
可以發現其實 Hierarchical Clustering 分類的結果跟 Original data 差不多





Agglomerative Clustering

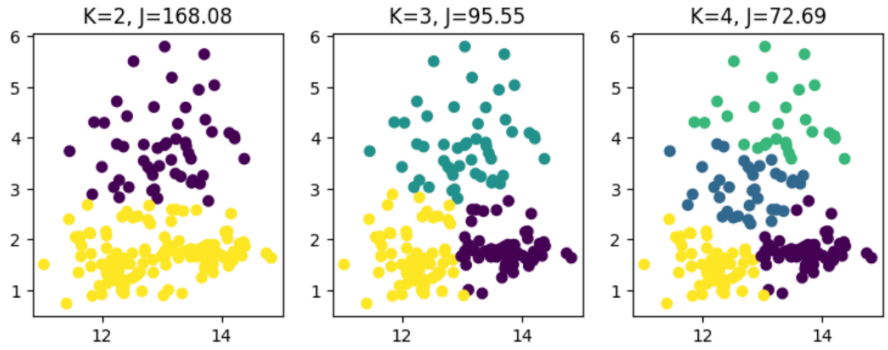
可以發現其實 Agglomerative Clustering 分類的結果跟 Original data 差不多



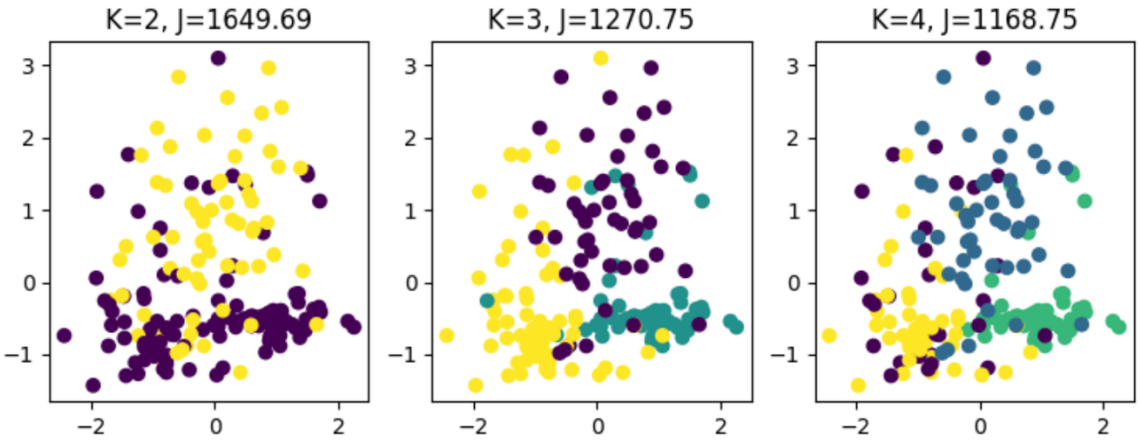
只看 Alcohol 和 Malicacid，可以發現它們分群分得還不錯，也就是代表在 13 個 feature 裡面有一些無用資訊會影響分群

	Alcohol	Malicacid
0	14.23	1.71
1	13.20	1.78
2	13.16	2.36
3	14.37	1.95
4	13.24	2.59
..	...	...
173	13.71	5.65
174	13.40	3.91
175	13.27	4.28
176	13.17	2.59
177	14.13	4.10

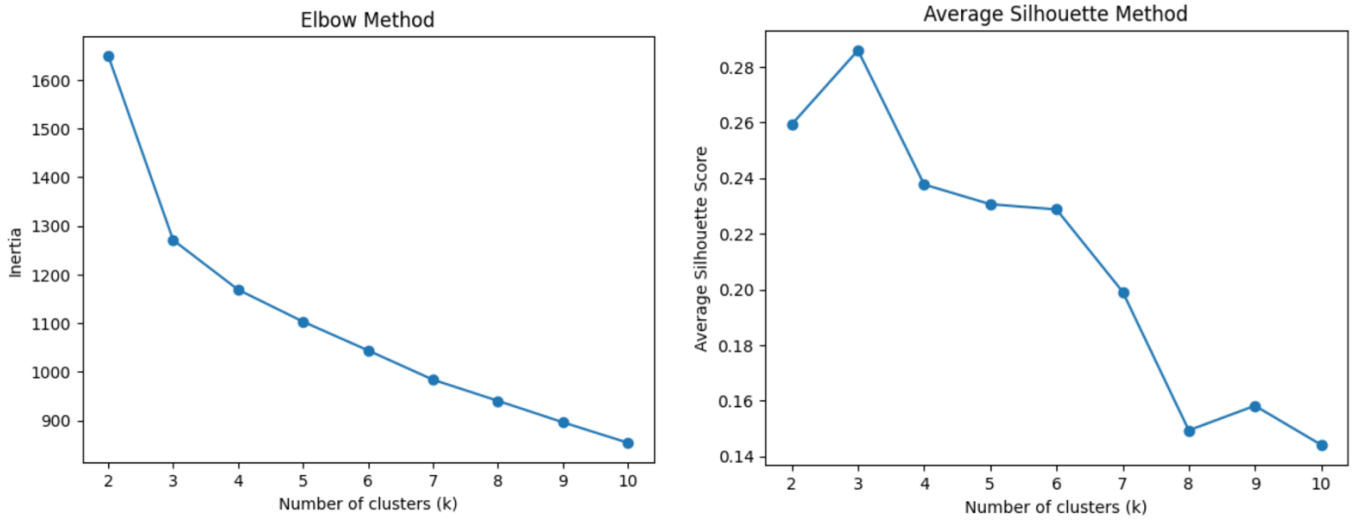
[178 rows x 2 columns]



Normalize 後的分群有分得稍微好一點



找最佳 K 值：Elbow Method / Average Silhouette Method

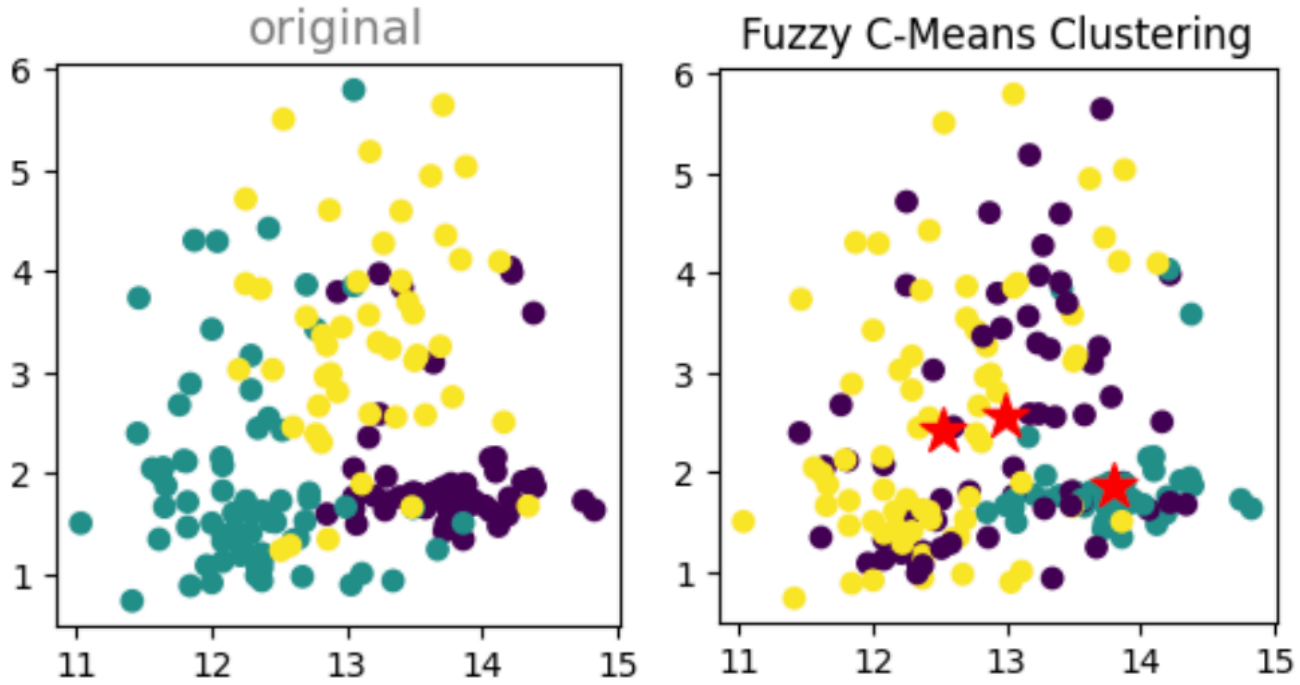


Elbow Method：隨著 k 值的增加，聚類的效果（比如集群內平方和）通常會迅速下降，直到某一個 k 值後下降的速率急劇減緩，形成類似彎肘的形狀。該「彎肘點」所對應的 k 值就是最佳的聚類數量，在這邊則是 3 的位置，所以當 k=3 時分類會分得最好

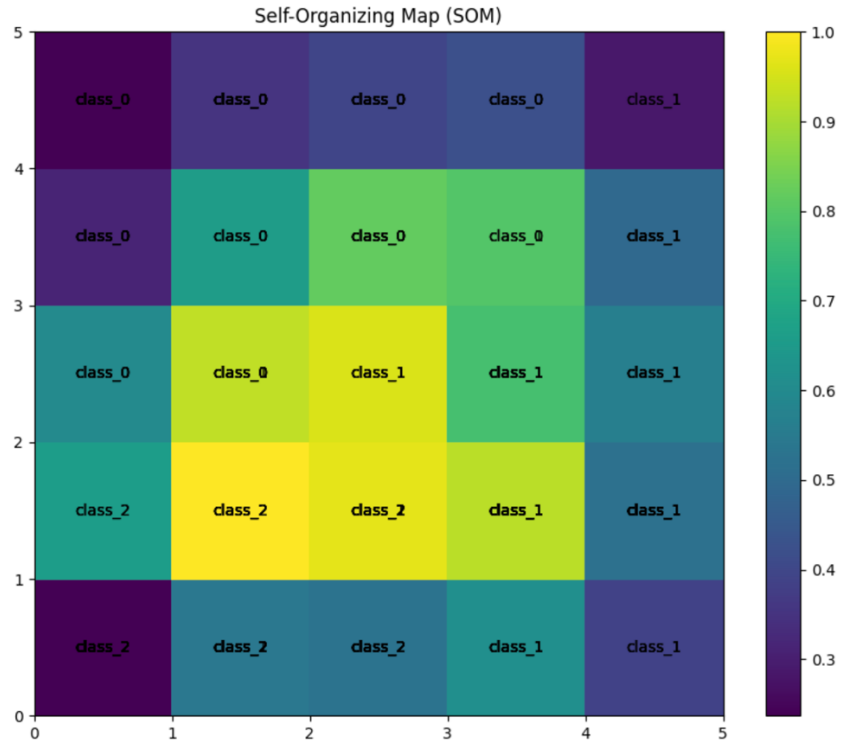
Average Silhouette Method：通過計算不同 k 值下的平均輪廓系數，來幫助選擇最佳的聚類數量。通常情況下，平均輪廓系數越接近 1，表示聚類效果越好，在這邊則是 3 的位置，所以當 k=3 時分類會分得最好

FCM

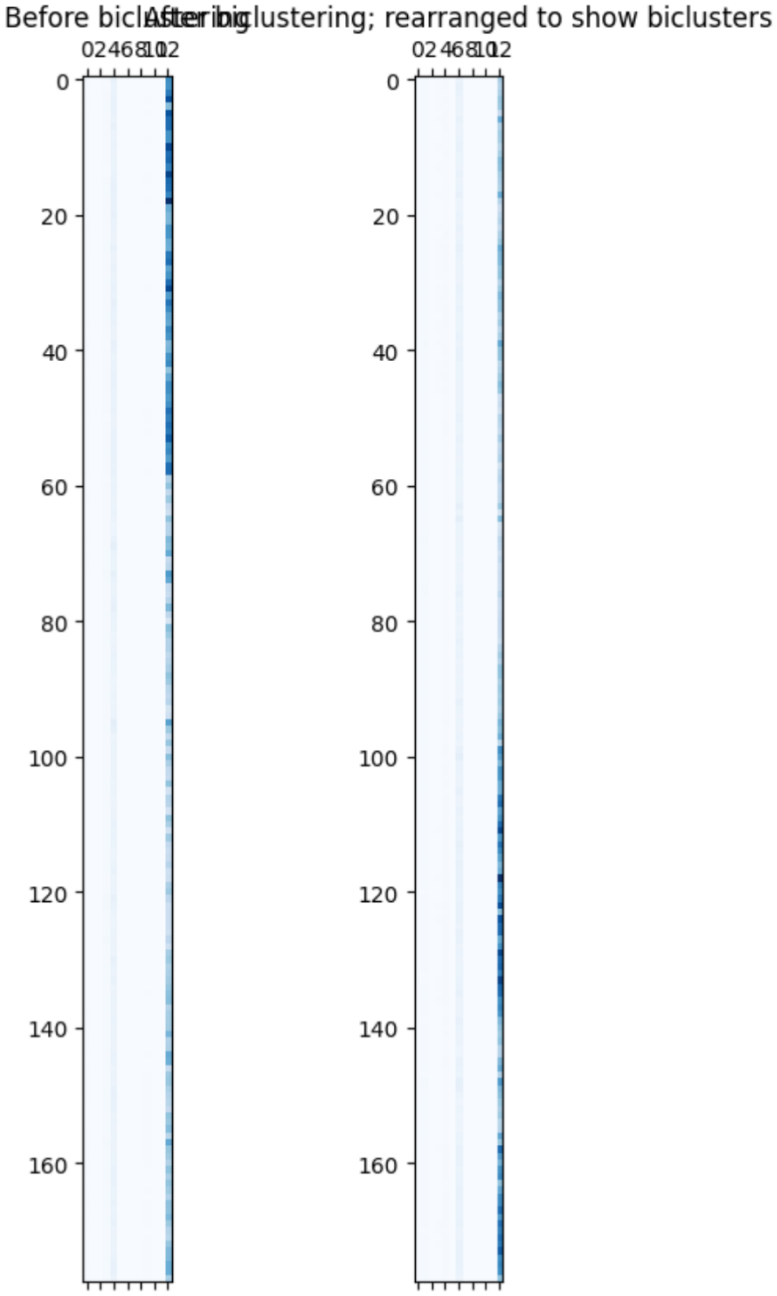
FCM (Fuzzy C-Means) 模型是一種基於模糊理論的聚類演算法，用於將資料點分組成不同的類別。它是 C-Means 聚類演算法的一種擴展，適用於那些具有模糊性質的資料集，即資料點可能屬於多個不同的類別，而不是嚴格地屬於某一個類別。



自組織映射（Self-Organizing Maps，SOM），也稱為 Kohonen 網絡，是一種用於將高維資料映射到低維空間的無監督學習神經網絡算法。它通常用於對高維資料進行聚類、可視化和特徵提取。



雙聚類（BiClustering）： 同時識別資料集中的行聚類和列聚類，從而找到資料集的子集，其中行和列具有類似的特徵



mean squared residue score (H) 是一種用於衡量聚類質量的指標，通常用於評估叢聚分析的性能

```python
from sklearn.datasets import load_wine
from sklearn.cluster import KMeans
import numpy as np

# wine
wine = load_wine()
X = wine.data

# 初始化KMeans模型并进行聚类
kmeans = KMeans(n_clusters=3, n_init=10, random_state=42)
kmeans.fit(X)

# 计算每个样本到其所属聚类中心的距离的平方
distances = np.min(kmeans.transform(X), axis=1)

# 计算平均平方残差分数 (H)
H = np.mean(distances)
print("Mean squared residue score (H):", H)
```

```
Mean squared residue score (H): 93.0094349215321
```