國立雲林科技大學

資訊管理研究所

資料探勘專案作業三

指導教授：　許中川　教授

學生：M11223038 陳品佑

M11223032 張祥恩

M11223033 鍾季衡

M11223036 魏冠宇

目錄

表目錄

圖目錄

Data Mining Homework: The Banana.csv, size3.csv

摘要

本研究分別使用 K-means, Hierarchical Clustering, DBSCAN，將 banana.csv 資料分成兩群，並計算分群所花費的時間，以及利用 Sum of Squares Error (SSE), Accuracy, Entropy 三個衡量指標來對三種演算法的分群效果做比較，並繪製出分類結果；另外也對 Size3.csv 資料集做以上的動作進行分群，最後以視覺化呈現出分群結果。而在 DBSCAN 的環境下，做了不同的參數設定來觀察對結果的影響。這也有助於評估 DBSCAN 對於不同資料及的適應性與效能。

關鍵字：K-means, Hierarchical Clustering, DBSCAN, Sum of Squares Error, Accuracy, Entropy.

Abstract

This study employs K-means, Hierarchical Clustering, and DBSCAN to partition the banana.csv dataset into two clusters. It measures the time taken for clustering and compares the clustering effectiveness of the three algorithms using Sum of Squares Error (SSE), Accuracy, and Entropy as evaluation metrics. The classification results are visualized, and a similar clustering analysis is performed on the Size3.csv dataset, with the outcomes presented visually. Additionally, various parameter settings are explored in the DBSCAN environment to observe their impact on the results. This aids in assessing the adaptability and performance of DBSCAN across different datasets.

Keywords：K-means, Hierarchical Clustering, DBSCAN, Sum of Squares Error, Accuracy, Entropy.

一、緒論

1.1 研究動機

群聚分析在資料科學上是一項重要的技術，能夠協助我們理解資料的特性以及分布結構和模式。而在實務上，K-means, Hierarchical Clustering 和 DBSCAN 等演算法皆是常見的分群演算法，但這些演算法都有各自的特性和限制。因此本研究旨在探討此三種演算法在不同資料及上的應用效果。

1.2 研究目的

本研究旨在深入探討 K-means, Hierarchical Clustering 和 DBSCAN 等群聚分析演算法在不同資料集上的應用效果，以及在 DBSCAN 環境下進行不同參數設定對分群結果的影響。具體目的包括：

1. **比較演算法效能**：透過對 Banana, Size3 資料集的群聚分析，評估 K-means, Hierarchical Clustering 和 DBSCAN 在分群效果上的差異。
2. **視覺化分類結果**：透過分群結果的視覺化呈現，有助於讓研究者或決策者更直覺地理解群聚分析的結果，並提高其分析的可解釋性。

二、研究方法

本研究於 Python 內分別先將 Banana.csv 以及 Size3csv 兩資料集匯入程式之後，即可開始針對 K-means, Hierarchical Clustering 和 DBSCAN 這三種演算法的分群實驗，同時分別製作可視化圖表以及效能、時間的分析與比較。

三、研究實驗

3.1 資料集簡介

本研究使用 Banana 與 Size3 做為本次研究主要資料集。Banana 資料集是一個在資料分析和機器學習領域常被使用的標準測試資料集之一。這個資料集通常被用來測試群聚分析演算法的性能，特別是在處理非線性和較複雜形狀的資料時。而 Size3 資料集做為一個在分群演算法中使用的資料集，本研究將利用此資料集來做各項演算法的效能評估以及分群表現。下表 1&2 為 Banana 資料集與 Size3 資料集之屬性簡介。

- Banana.csv
- 資料集名稱：Banana
- 資料筆數：4811
- 屬性數量：3 (with label)

表 1　Banana 資料集屬性簡介

屬性名稱	型態	尺度
X	Int	Nominal
Y	Int	Nominal
Class	Int	Ordinal

- Size3.csv
- 資料集名稱：Size3
- 資料筆數：1000
- 屬性數量：3 (with label)

表 2　Size3 資料集屬性簡介

屬性名稱	型態	尺度
X	Int	Nominal
Y	Int	Nominal
Class	Int	Ordinal

3.2 前置處理

本研究在進行分析前，首先載入需要的演算法套件，以及檢查資料品質，例如：檢查資料中是否含有缺失值、空值等等。

3.3 實驗設計

在 Python 上使用 K-means, Hierarchical Clustering 和 DBSCAN 演算法對 Banana 資料集以及 Size3 資料集做分群，並以(1). Sum of Squares Error (SSE) (2). Accuracy (3). Entropy 三項績效指標呈現並分析其效能。最後針對 DNSCAN 演算法做不同參數的調整來比較分群結果。

3.4 實驗結果

3.4.1 Banana 資料集 (K-means)

以下分別對各個演算法在 Banana 資料集的分群結果呈現，分析並比較其效果。可以看出 K-means 在此資料集上的效果能夠大致上正確地將這兩群分類。如下圖 1。

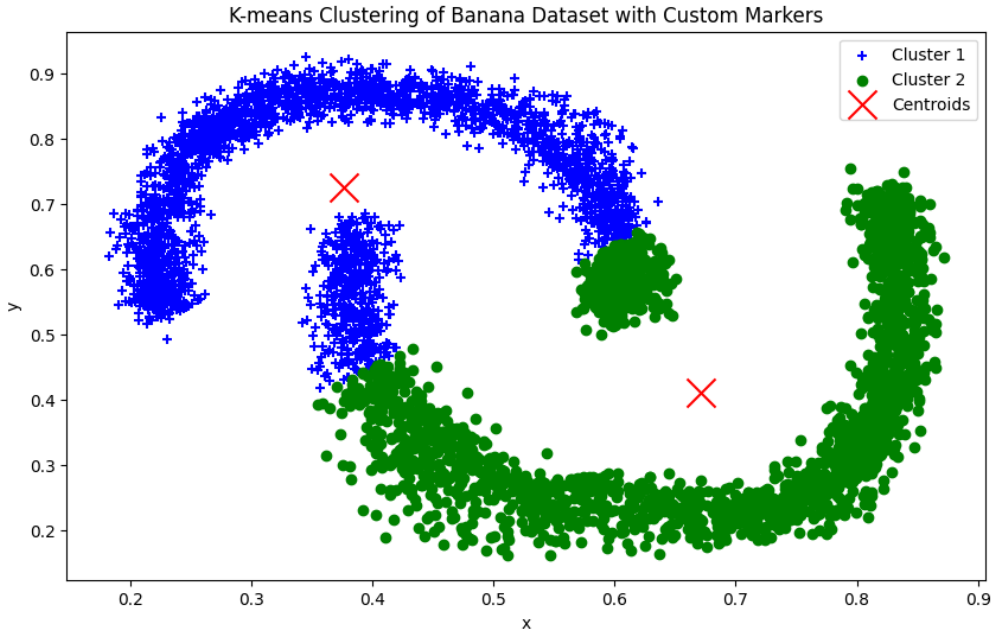

圖 1 K-means 散點圖

同時本研究也將分群的數量與所產生的平方誤差總和(SSE)繪製出來，使得更容易看出在分的群數越多時，SSE 會帶來什麼樣的改變。如下圖 2。

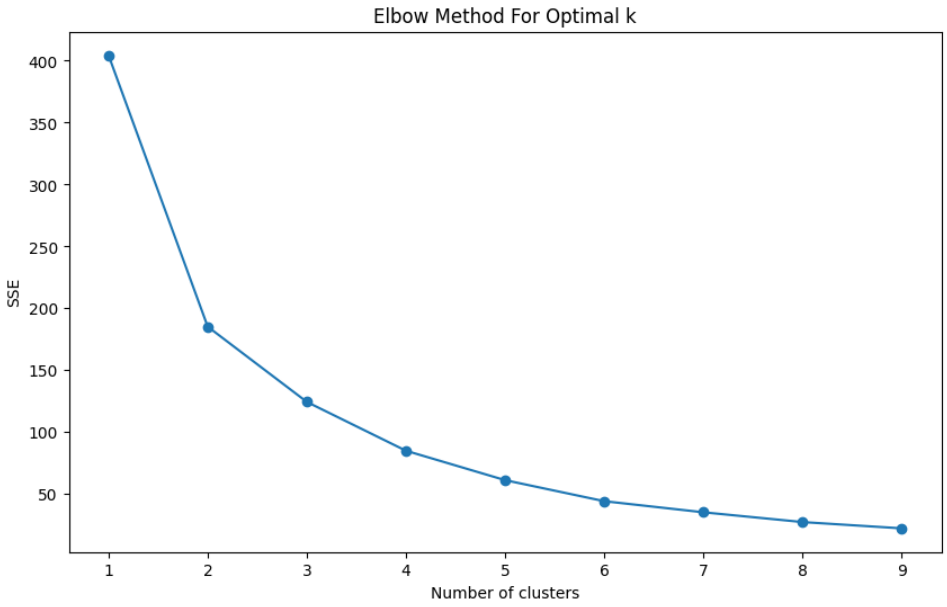

圖 2　Elbow Method For Optimal "K" - Banana

3.4.2 Banana 資料集 (Hierarchical Clustering)

在 Hierarchical Clustering 階層式分群法中，本組繪製出可視化圖形，用以呈現分群結果概況如下圖 3。而在此分群演算法相較於 K-means 的分群效果較差，本組推測可能原因為階層式分群法對於非線性結構資料分佈狀況更加敏感。因此 K-means 的表現會優於階層式分群。如下圖 4。

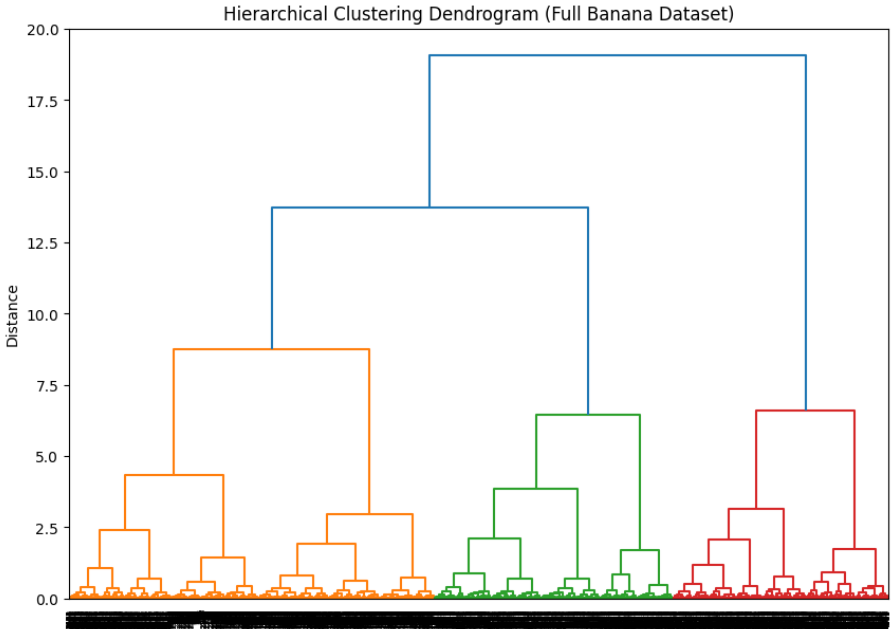

圖 3　Hierarchical Clustering of Banana Dataset

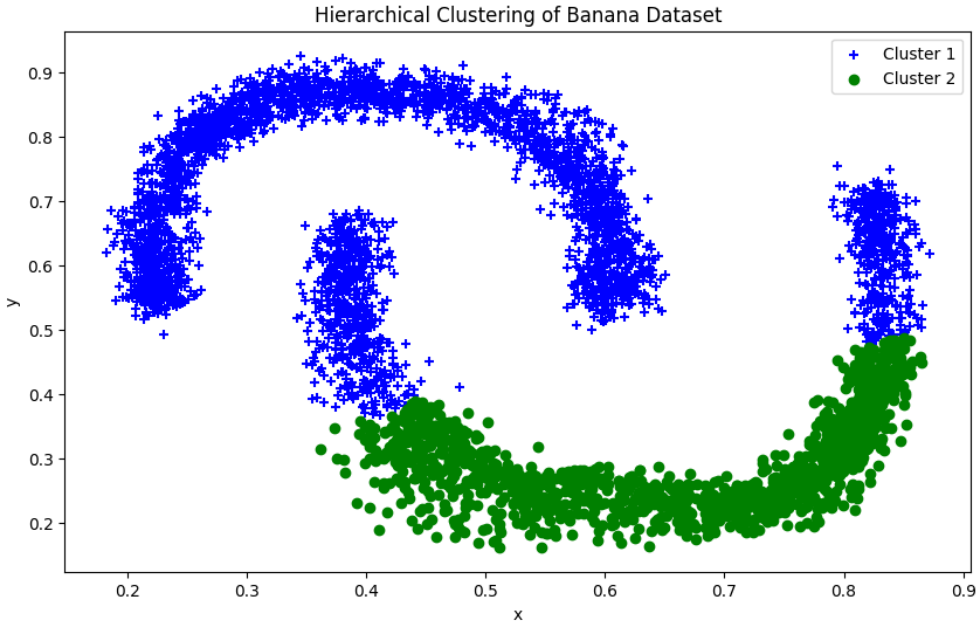

圖 4 Hierarchical Clustering of Banana Dataset

3.4.3 Banana 資料集 (DBSCAN)

最後在 Banana 資料集上使用 DBSCAN 演算法做分群。而得出的效果顯示，相較於先前兩個演算法的分群效果，DBSCAN 的準確度表現最為優異如圖 5 所示。而本組也利用不同的視覺化圖形呈現資料分佈的狀況。圖 6 顯示的是以矩陣形式呈現資料點之間的距離。圖 7 則以 3D 可視化資料分佈情形。

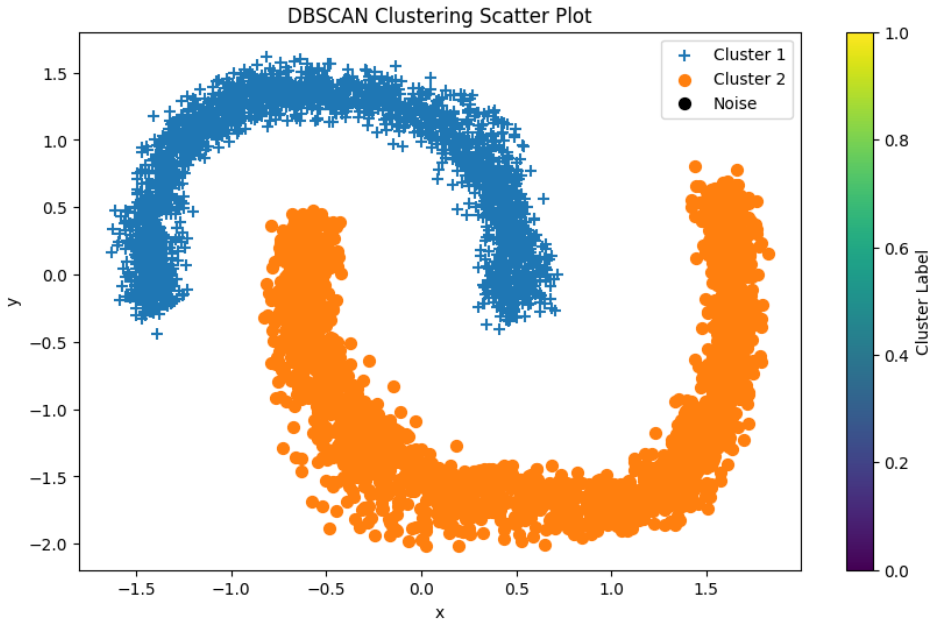

圖 5 DBSCAN Clustering Scatter Plot

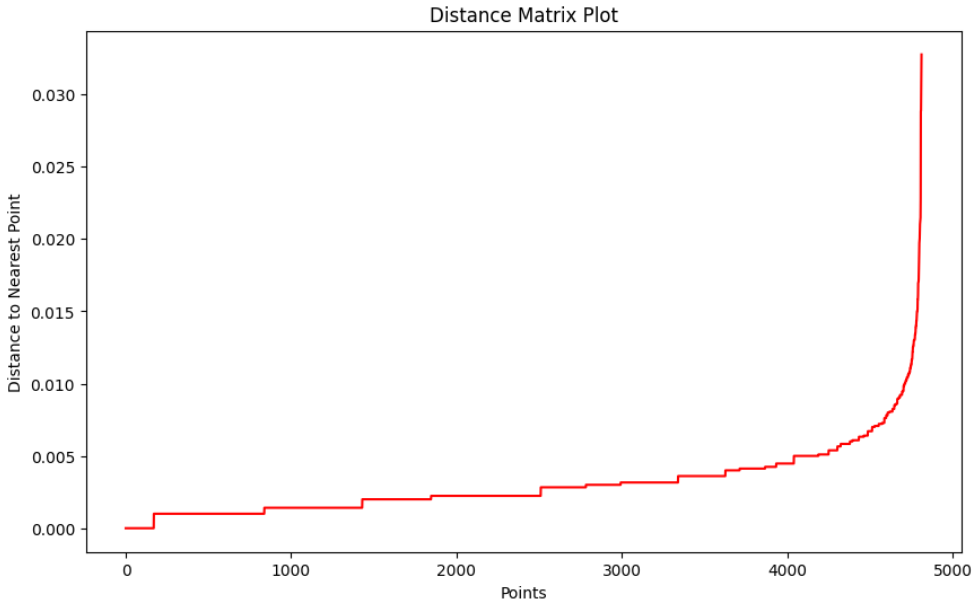

圖 6 Distance Matrix Plot - Banana

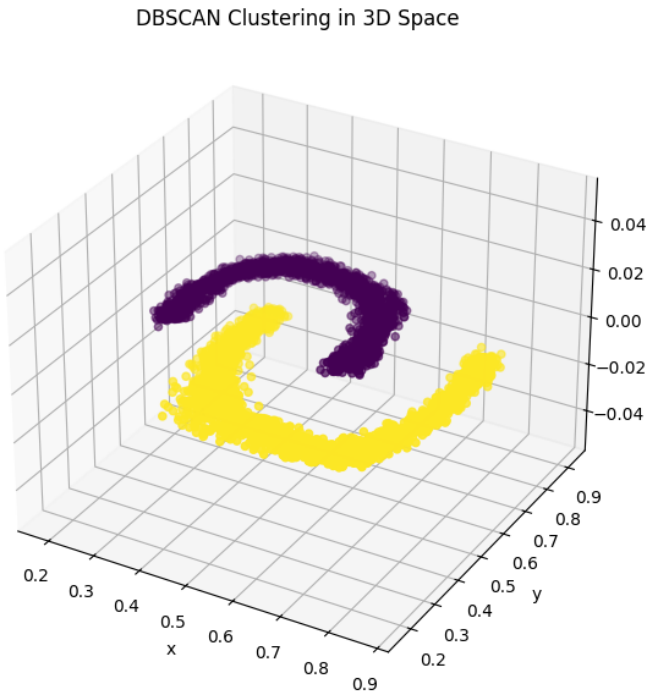

圖 7 3D Space - Banana

3.4.4 Size3 資料集 (K-means)

接著，本組以 Size3 資料集實作 K-means 群聚演算法，並分析其效能表現。本研究將其資料分為四群如下圖 7。K-means 演算法對於 Sizes3 資料集的準確度來到 98%，效果相當優異。而下圖 9 為

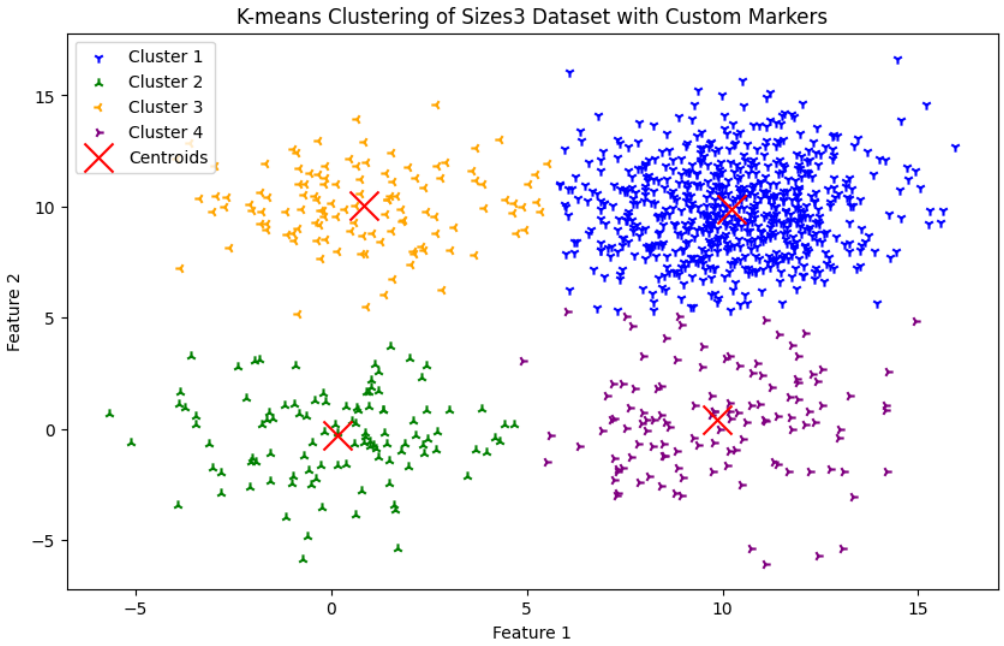

圖 8 K-means Clustering of Sizes3 Dataset

本研究一樣也將分群的數量與所產生的平方誤差總和(SSE)繪製出來，使其更容易看出在分的群數越多時，SSE 會帶來什麼樣的改變。如下圖 9。

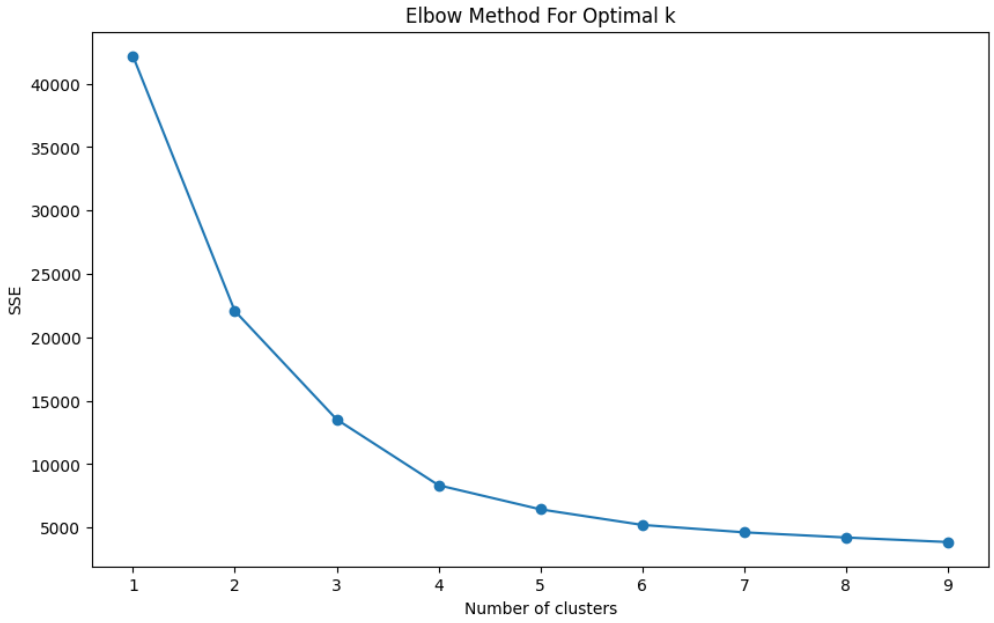

圖 9 Elbow Method For Optimal "K" - Size3

3.4.5 Size3 資料集 (Hierarchical Clustering)

本組也對其資料集實作出階層式分群演算法的分析。同樣地將本資料集分作四群，由實驗結果得知，階層式分群的準確度為 98.6%，其效能與 K-means 差異不大。由下圖 10&11 可得知階層圖以集資料分佈狀況。

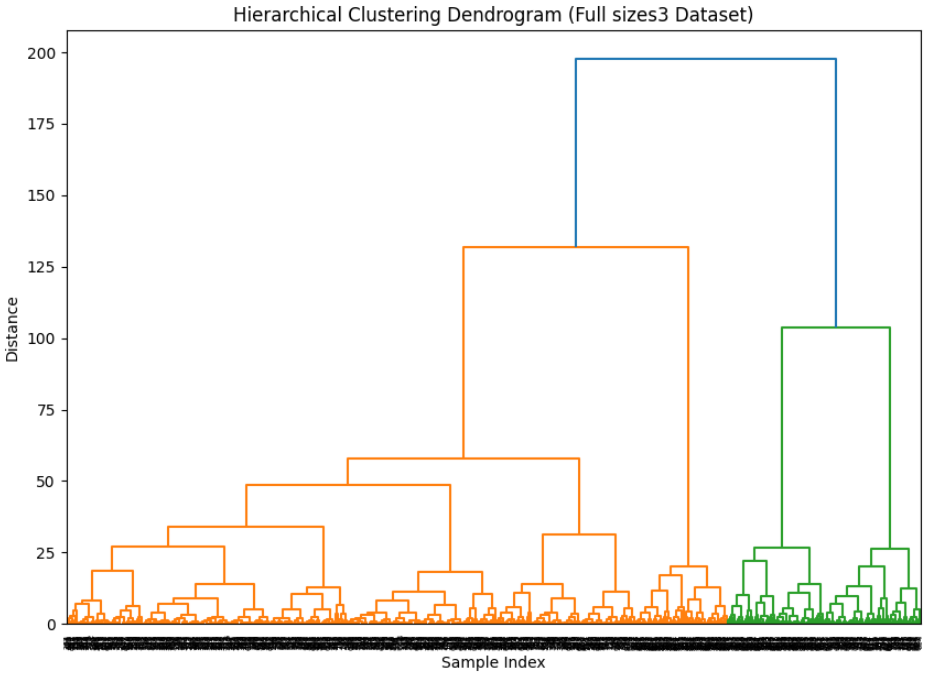

圖 10 Hierarchical Clustering Dendrogram (Full sizes3 Dataset)

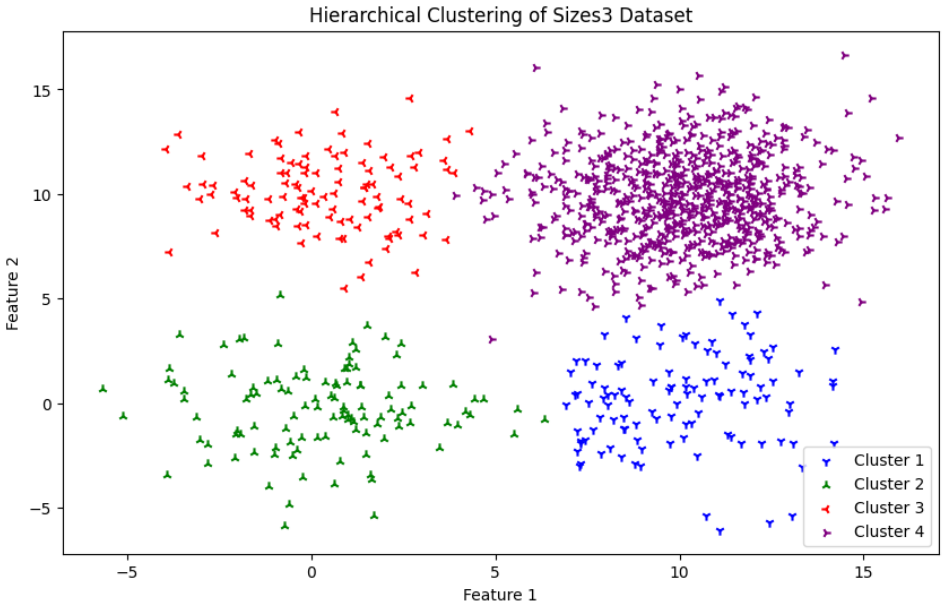

圖 11 Hierarchical Clustering of Sizes3 Dataset

3.4.6 Size3 資料集 (DBSCAN)

最後則是實作 DBSCAN 於 Size3 資料集上的分群表現，在此演算法上，本研究分別在參數調整上面分別作分兩群、三群、以及四群的參數，以此觀察效果的變化。如下圖 12, 13&14。在分群準確率上，分做四群的表現最優異為 94.1%，可想而知分做兩群的效果十分不好為 67.1%。

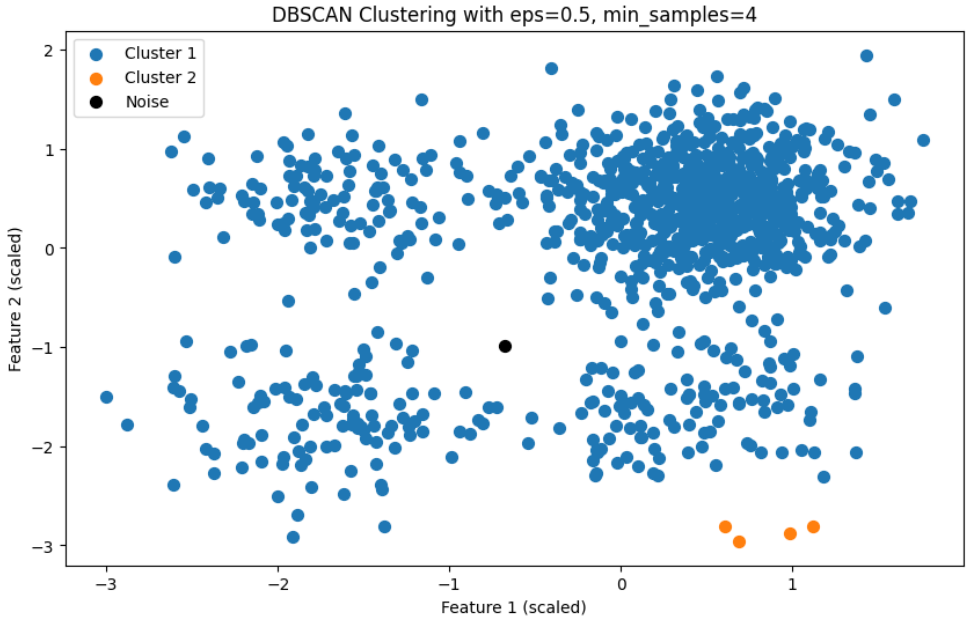

圖 12 DBSCAN Clustering with eps=0.5, min_samples=4

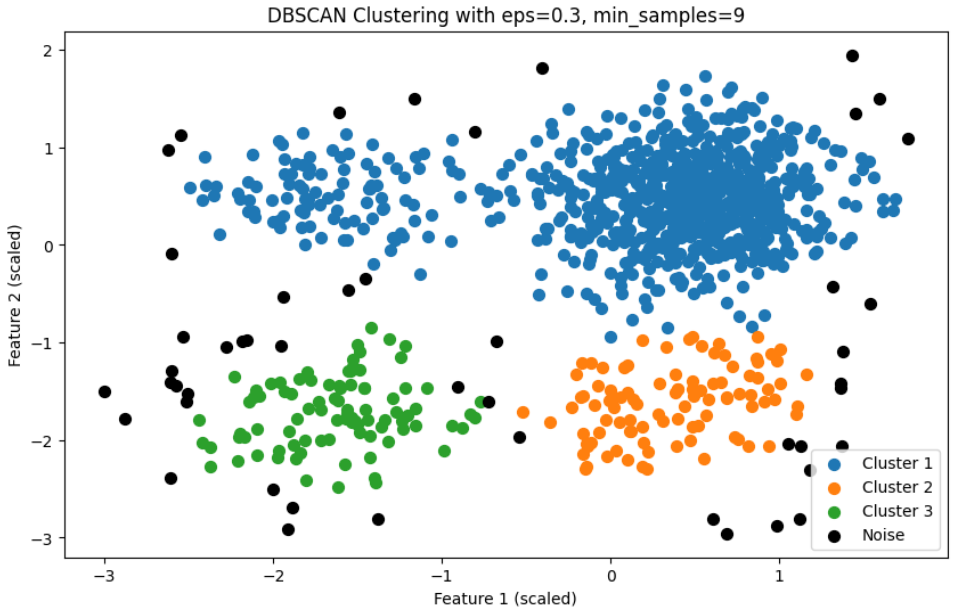

圖 13 DBSCAN Clustering with eps=0.3, min_samples=9

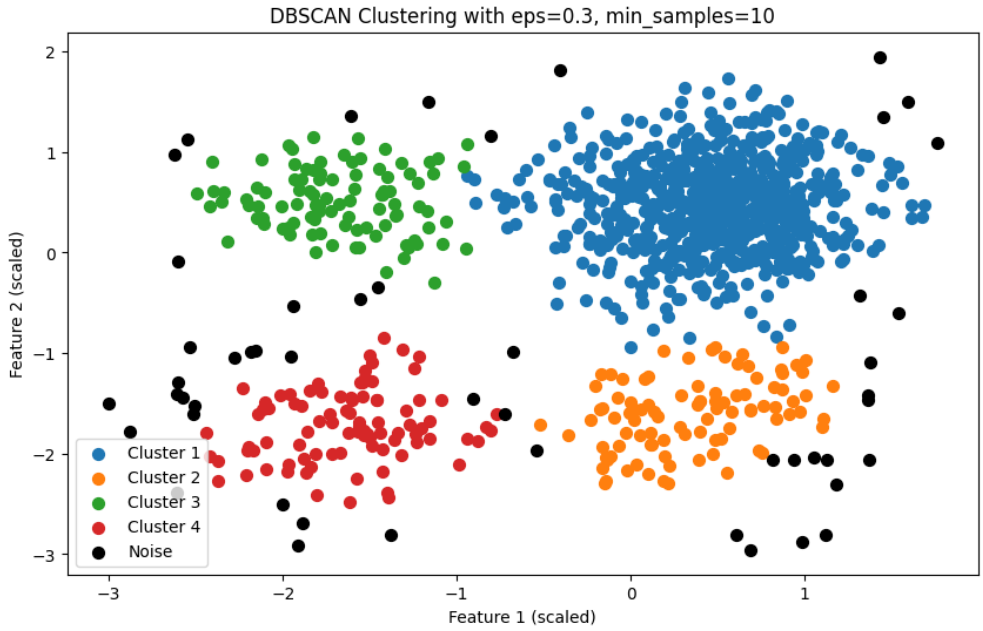

圖 14 DBSCAN Clustering with eps=0.3, min_samples=10

本實驗同樣為資料集做 3D 空間的資料分佈狀態以及在距離矩陣上的視覺化呈現，如下圖 15&16 所示。

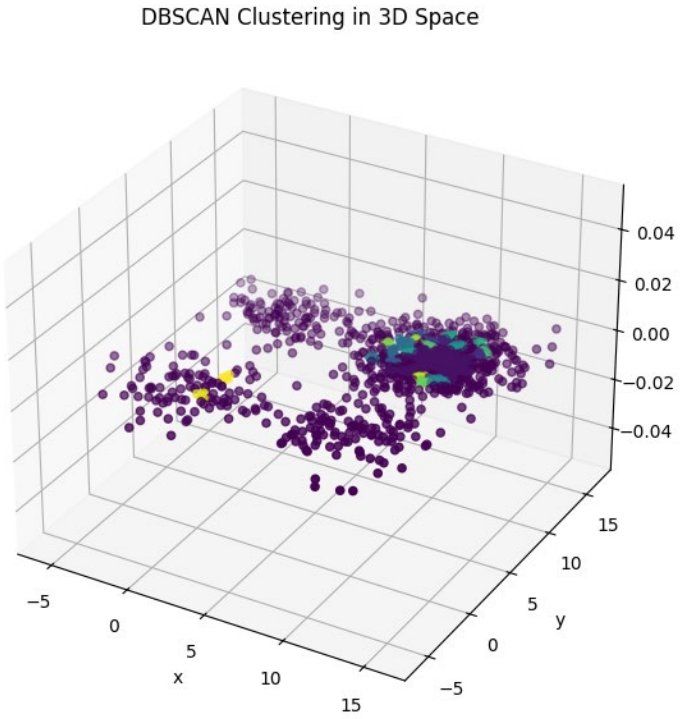

圖 15 3D Space - Size3

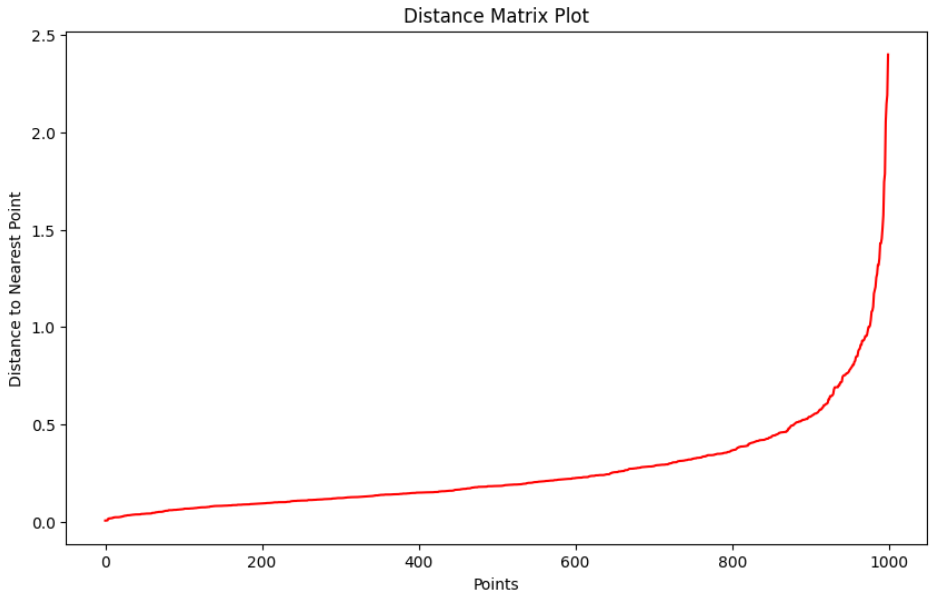

圖 16 Distance Matrix Plot - Size3

由以上的實驗結果可製作出 Banana 資料集和 Size3 資料集對應至各演算法之間的效能比較表如下表 3&4，

表 3 Banana 之三種演算法績效

Banana 績效	K-means	Hierarchical Clustering	DBSCAN
SSE	185.20	222.59	227.12
Accuracy	0.83	0.79	1.00
Entropy	0.46	0.30	0.00
Time(second)	3.95	31.07	7.67

表 4 Size3 之三種演算法績效

Size3 績效	K-means	Hierarchical Clustering	DBSCAN
SSE	8341.25	2689.19	349.00
Accuracy	0.98	0.986	0.94
Entropy	0.15	0.086	1.34
Time(second)	1.28	3.69	2.25

四、結論

本研究基於對 K-means、Hierarchical Clustering 和 DBSCAN 在 Banana dataset 和 Size dataset 上進行的實驗可得出在對於分線性資料的處理上，使用 K-means 進行分群時，由於其對於資料點分佈的形狀較其他演算法敏感，導致分群的效果表現較差。相對的，DBSCAN 在處理形狀較為特殊的表現較為優異。然而在程式的計算時間表現上來說，Hierarchical Clustering 階層式分群的時間表現是最差的，由此也驗證了本課程老師所提及之階層式分群在資料量龐大之時，其時間複雜度將可能來到$O(n^2)$。而對於不同群數的處理上，K-means 與階層式分群在處理 Size3 資料集的四個群數表現較為穩定，由此也反映了它們對於均勻分布的群聚結構較具有適應性，相較之下，DBSCAN 在處理不同群數時受到參數設定上的影響，需要更加細微的調整才有辦法使其表現能夠更佳。

五、參考文獻

Abid Ali Awan(2022/8/17). Implementing DBSCAN in Python. https://www.kdnuggets.com/2022/08/implementing-dbscan-python.html

Comparing different clustering algorithms on toy datasets. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py

sklearn.cluster.DBSCAN. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

Trying to plot outliers using DBSCAN. https://stackoverflow.com/questions/62785623/trying-to-plot-outliers-using-dbscan

Stack overflow(2019)。Clustering the 3D points when given the x,y,z coordinates using DBSCAN algorithm using python。Clustering the 3D points when given the x,y,z coordinates using DBSCAN algorithm using python

Sushant Kafle (RIT Student)(2018). DBSCAN。https://github.com/SushantKafle/DBSCAN/blob/master/dbscanner.py

Eric Plog(2018/7/10). Functions to Plot KMeans, Hierarchical and DBSCAN Clustering. https://medium.com/@plog397/functions-to-plot-kmeans-hierarchical-and-dbscan-clustering-c4146ed69744

Michael Fuchs (2020/6/15). DBSCAN. https://michael-fuchs-python.netlify.app/2020/06/15/dbscan/

趙孝正(2022/8/22)。聚類演算法評價指標 Adjusted Rand Index, ARI 指数。https://blog.csdn.net/weixin_46713695/article/details/126388445