資料探勘專案作業三

利用 Python 軟體實作群聚分析

指導教授：

許中川 教授

成員：

M11112075 徐紹鈞

M11121028 陳韋諠

日期：

2022 年 12 月 15 日

摘要

本研究利用了 Iris 以及 Bank 資料集做集群分析(Cluster Analysis)，是一種統計方法，常應用在語言分析及電腦科學領域上，用途相當的廣泛，本研究將會利用 K-means、DBSCAN 及階層式分群進行分群，並比較三種方法的時間、純度、calinski_harabasz_score，可以發現 K-means 花費在分群的時間上較其兩種方法短，故 k-means 為兩個資料集最佳的分群方法。

一、 緒論

1.1 動機

1.1.1 Iris Dataset

花卉一直是讓台灣被全世界看到的原因之一，眾多的花卉種類也讓台灣舉辦了許多的花博活動，也吸引了許多國外的人前來觀看，若可以利用分群將相似高的花卉分群，在調整擺放花卉的位置時，就可以將相似度高的花卉放在一起，讓遊客可以流暢的暢遊花博。

1.1.2 Bank Dataset

在金融相關產業中為了尋求在直接營銷活動中客戶是否會認購定期存款，若可以利用客戶基本資料、當前活動的最後一次聯絡人有關資料以及社會和經濟背景屬性來分群，變可以為同群的民眾設置購買策略，以利於利潤最大化。

1.2 研究目的

1.2.1 Iris Dataset

目前的演算法有相當多種，本研究利用 K-means、DBSCAN 以及階層式分群三種來進行分群，投入萼片長度、萼片寬度等 4 種屬性資料，希望透過三種分群方式找到最適合的分群方式，並投入實際情況來做使用

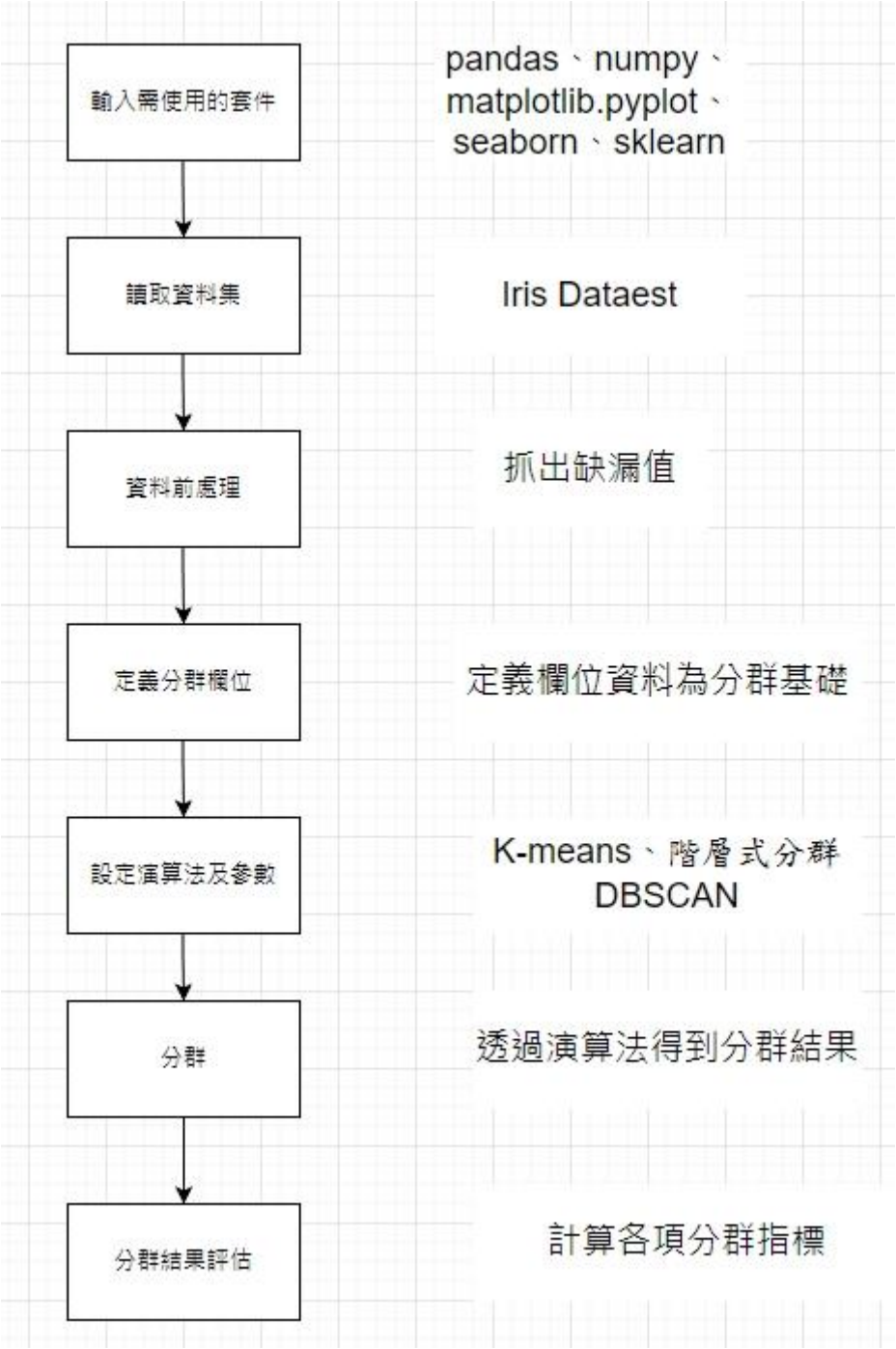
1.2.2 Bank Dataset

為了為銀行抓取重要客戶，本研究將會投入 Bank 資料集中的年齡、工作、教育程度等 16 種屬性，利用 K-means、DBSCAN 以及階層式分群三種來進行分群，希望透過三種分群方式找到最適合的分群方式，並投入實際情況做使用

二、 方法

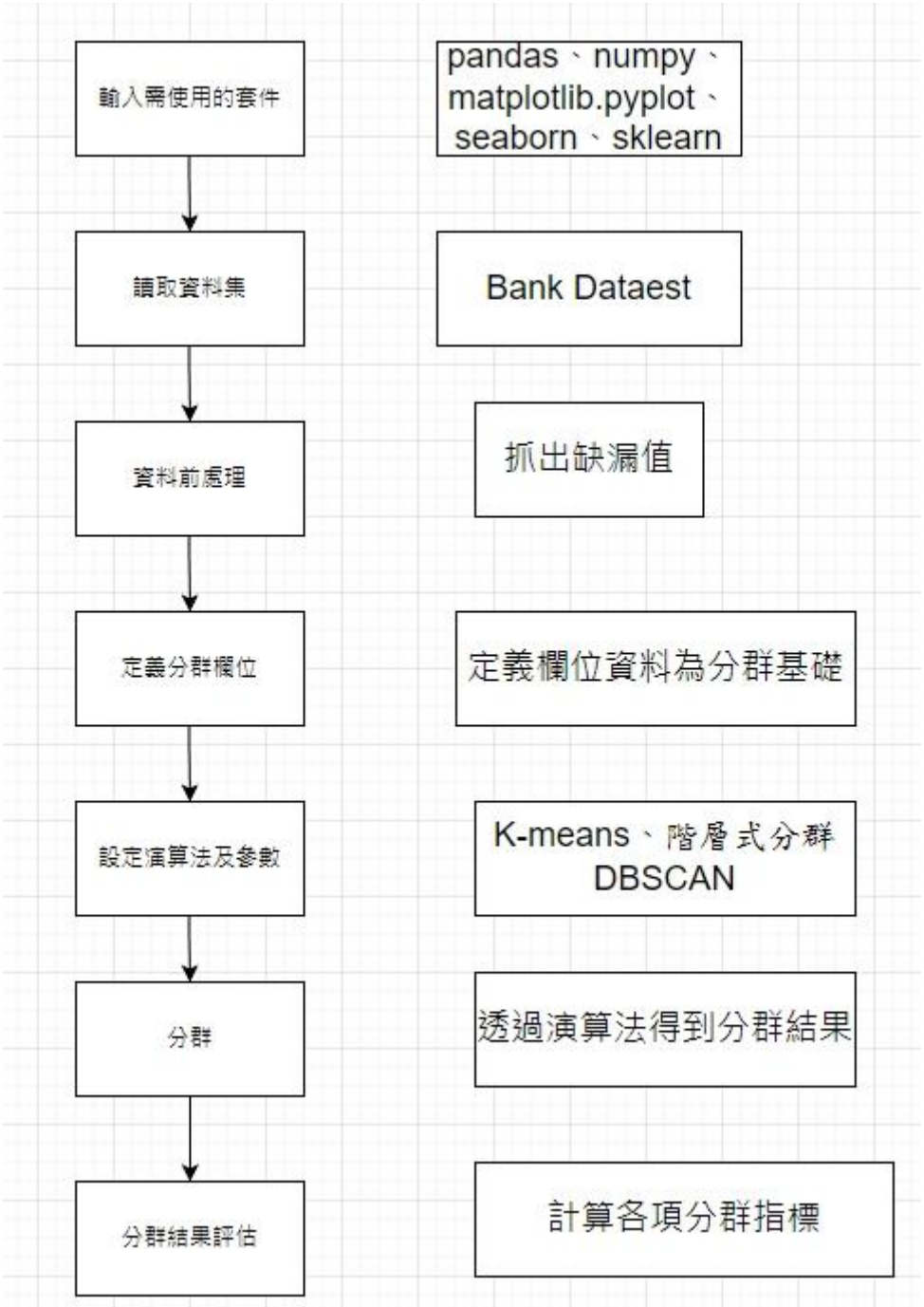
2.1 程式架構

2.1.1 Iris Dataset



圖一 Iris Dataset 之程式架構流程圖及說明

2.1.2 Bank Dataset



圖二 Bank Dataset 之程式架構流程圖及說明

2.2 程式執行方法

2.2.1 **K-means**

k-means 是目前廣泛運用的分群演算法之一，可使用在龐大的資料集上，此分群法是依據資料的類似程度去歸類，同時 k-means 也是一種簡單的非監督式學習演算法，可將資料分為給定的群數。

k-means 演算法流程如下:

步驟一:首先將會決定給定的群數，設定群數為 K 群，接著將會從樣本中挑出 K 個樣本作為群聚中心。

步驟二:將會輸入全部的樣本，計算每筆樣本對每個群聚中心的距離，接著比較該筆資料對哪個群集中心較接近，這筆資料將會被納入距離最近的群集中心。

步驟三:根據群內的每一個樣本計算該群集的質量中心，利用新的質量中心當作新的群集中心，接著在比較每一筆資料與新的群集中心的距離，根據距離，再重新分配每一筆資料所在的群集。

2.2.2 DBSCAN（**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise）

DBSCAN（**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise）是一種非監督式學習，是一種基於密度來進行聚類的演算法，會將特徵相近且密度高的樣本劃分為一群，故此方法的優點是不受極端質影響，因為利用密度分群，故極端值會自成一群。

2.2.3 階層式分群（**hierarchical clustering)**

階層式分群（**hierarchical clustering)**，此方法是透過階層架構的方式，將資料一層層的反覆進行分裂或者聚合，並產生最後的樹狀圖，此方法的優點有簡單明瞭，建構完完整的樹狀分類，可以方便的決定要分成幾群，缺點則是此方法僅適用於小樣本的資料。

三、 實驗

3.1 資料集

3.1.1 Iris Dataset 說明

此資料集建立於 1998 年 7 月 1 日共有 150 筆，5 個欄位。

表一 Iris Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
sepal length in cm	萼片長度 cm
sepal width in cm	萼片寬度 cm
petal length in cm	花瓣長度 cm
petal width in cm	花瓣寬度 cm
Iris Setosa	鳶尾
Iris Versicolour	雜色鳶尾
Iris Virginica	弗吉尼亞鳶尾

3.1.2 Bank Dataset 說明

此資料集建立於 2018 年 9 月 7 日共有 4899 筆，17 個欄位。

表二 Bank Dataset 欄位資料說明彙總表

age	年齡
job	工作
marital	婚姻
education	教育
default	NA
housing	房產
loan	貸款
contact	合約
month	月份
day_of_week	星期
duration	期間
campaign	活動
pdays	工作日
previous	先前
poutcome	結果
emp.var.rate	就業變動率
cons.price.idx	消費者價格指數
cons.conf.idx	消費者信心指數

3.1.3 實驗數據

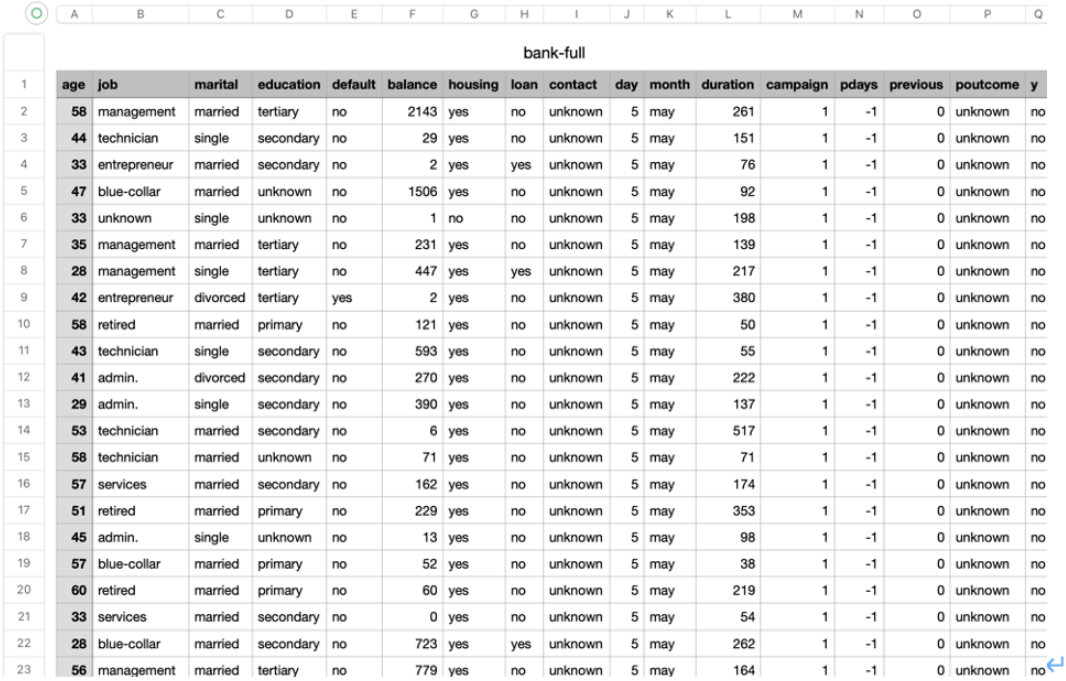
(1) Iris Dataset 資料集

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.1	3.5	1.4	0.2	0.0
1	4.9	3.0	1.4	0.2	0.0
2	4.7	3.2	1.3	0.2	0.0
3	4.6	3.1	1.5	0.2	0.0
4	5.0	3.6	1.4	0.2	0.0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2.0
146	6.3	2.5	5.0	1.9	2.0
147	6.5	3.0	5.2	2.0	2.0
148	6.2	3.4	5.4	2.3	2.0
149	5.9	3.0	5.1	1.8	2.0

150 rows × 5 columns

圖三 Iris Dataset 資料集

(2) Bank Dataset 資料集



bank-full

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no
28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no
42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no
58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no
43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no
41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no
29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no
53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no
58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no
57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no
51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no
45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no
57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no
60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no
33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no
28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no
56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no

圖一

圖四 Bank Dataset 資料集

3.2 前置處理

3.2.1 Iris Dataset

將 Iris 資料集，接著在觀察檔案時發現有許多資料有缺值的現象產生，我們將有缺值的資料進行刪除，完成資料清洗，最後載入進 jupyter Notebook。

3.2.2 Bank Dataset

將 Bank 資料集，接著在觀察檔案時發現有許多資料有缺值的現象產生，我們將有缺值的資料進行刪除，完成資料清洗，最後載入進 jupyter Notebook。

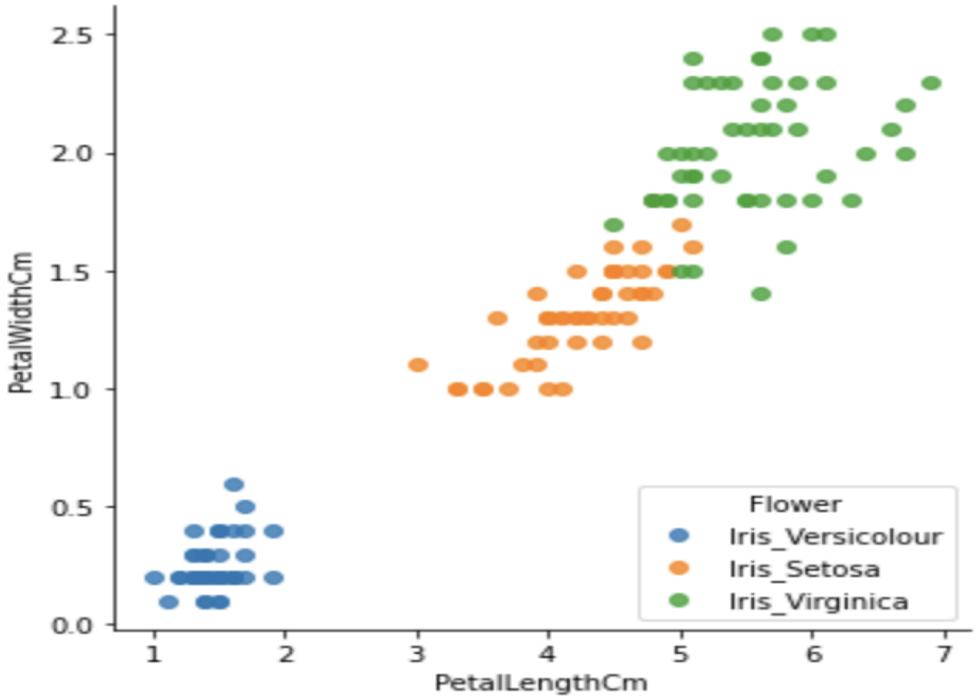
3.3 實驗設計

3.3.1 Iris Dataset

(1)套件載入：利用 Sklearn 載入 K-means 分群套件、階層式分群套件(AgglomerativeClustering)以及 DBSCAN 分群套件以利後續分群，接著載入繪圖套件(matplotlib)，以利後續指標績效評估

(2)定義分群基礎:根據 Label 欄位的資料為分群基礎(Iris Setosa =0, Iris Versicolour =1, Iris Virginica=2)

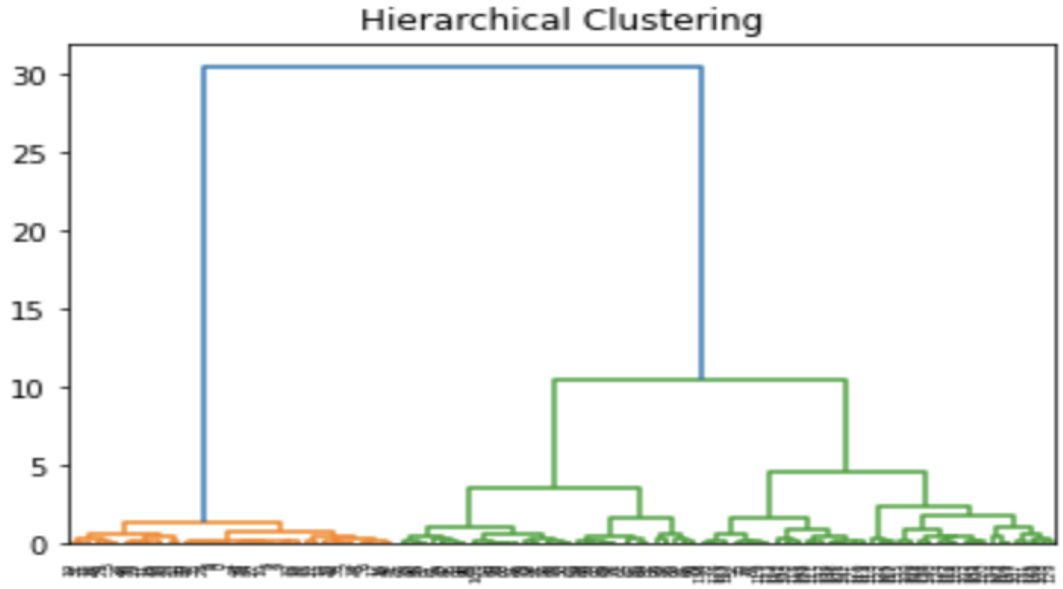
(3)K-means：將資料分成 3 類，利用 from sklearn.cluster import KMeans 來查看 K-means 分群結果如下圖(五)所示以及 Purity 指標圖(六)。




```
#Purity指標衡量
import numpy as np
from sklearn import metrics
def purity_score(y_true, clusters_pred):
    contingency_matrix = metrics.cluster.contingency_matrix(y_true, clusters_pred)
    return np.sum(np.amax(contingency_matrix, axis=0)) / np.sum(contingency_matrix)
purity_score(y_true, clusters_pred)
```

```
0.8933333333333333
```

(4)階層式分群：將資料分成 3 群，距離的計算方式使用 euclidean，群與群之間的距離使用 ward，最後畫出劃出階層式分群的階層樹(Dendrogram)，圖(七)及 Purity 指標圖(八)。




```
import numpy as np
from sklearn import metrics
def purity_score(y_true, y_pred):
    contingency_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    return np.sum(np.amax(contingency_matrix, axis=0)) / np.sum(contingency_matrix)
```

```
purity_score(y_true, y_pred)
```

```
0.96
```

圖(六)

(5)DBSCAN：將 eps 設成 0.7, min_samples 設成 3，並比較原始分群與使用 DBSCAN 的差異圖(九)及 Purity 指標圖(10)。

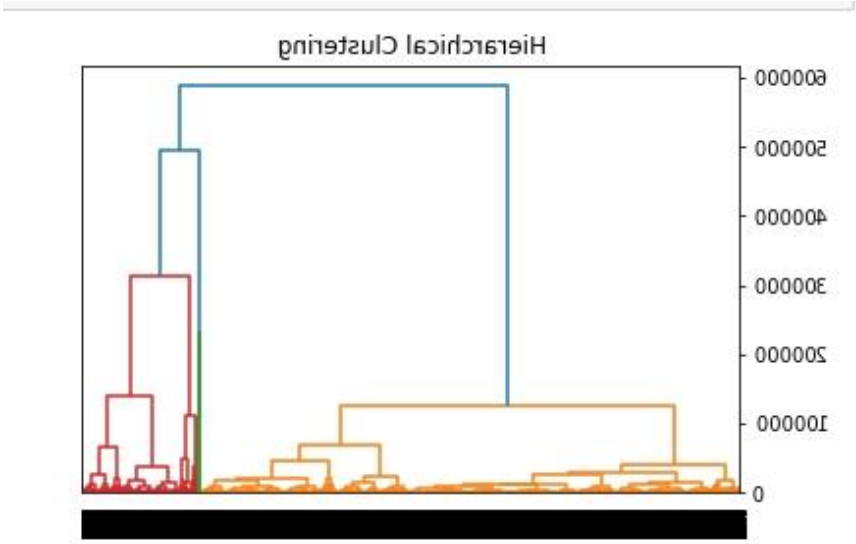
3.3.2 Bank Dataset

(1)套件載入：利用 Sklearn 載入 K-means 分群套件、階層式分群套件(AgglomerativeClustering)以及 DBSCAN 分群套件以利後續分群，接著載入繪圖套件(matplotlib)、calinski_harabasz_score 套件，以利後續指標績效評估。

(2)定義分群基礎:根據 Label 欄位的資料為分群基礎(no=0,yes=1)

(3)K-means：將 k-means 的參數設定設為 3 群，接著利用 calinski_harabasz_score 及純度進行績效評估。

(4) 階層式分群：將資料分成 5 群，距離的計算方式使用 euclidean，群與群之間的距離使用 ward，最後畫出劃出階層式分群的階層樹



(5)DBSCAN：將 DBSCAN 方法的參數設定 eps 設成 5, min_samples 設成 3

3.4 實驗結果

3.4.1 Iris 績效評估

	K-means	階層式分群	DBSCAN
分群花費時間(s)	0	0	0.1
Purity	0.89	0.96	0.67

表三 Iris 績效評估

3.4.2 Bank 績效評估

	K-means	階層式分群	DBSCAN
分群花費時間(s)	4	180	30
Purity	0.883	0.883	0.883
calinski_harabasz_score	61854	81969	0.90

表四 Bank 績效評估

四、 結論

Iris 資料集在 k-means 及階層式分群有相同的時間，但 Purity 是階層式分群來的高，故階層式分群的績效最好，其次則為 K-means。

Bank Dataset 資料集若要以時間作為評估的話，最佳為 k-means，其次為 DBSCAN

若要以 calinski_harabasz_score 作為評估，最佳為階層式分群，其次為 K-means。

五、 參考文獻

[1] Iris 資料集：https://archive.ics.uci.edu/ml/datasets/iris

[2]K-means 分群: https://ithelp.ithome.com.tw/articles/10209058

[3]階層式分群：http://mirlab.org/jang/books/dcpr/dcHierClustering.asp?title=3-2%20Hierarchical%20Clustering%20(%B6%A5%BCh%A6%A1%A4%C0%B8s%AAk)&language=chinese

[4]階層式分群：https://jamleecute.web.app/hierarchical-clustering-階層式分群/

[5]淺談階層式分群法：https://ithelp.ithome.com.tw/articles/10296825?sc=iThelpR

[6]基於密度的聚類演算法 DBSCAN: https://jason-chen-1992.weebly.com/home/-dbscan

[7]建立多個子圖表（ subplot、subplots): https://steam.oxxostudio.tw/category/python/example/matplotlib-subplot.html

[8] bank Dataset: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing