摘要

本實驗基於 UCI datasets 中的 Iris Data Set 資料集以及 Abalone Data Set 資料集作為研究樣本，前者為 1936 年，對加斯帕半島上的不同亞屬鳶尾花所提取的花瓣及花萼之長寬資料，後者 1994 年，是塔斯馬尼亞鮑魚的種群生物學，來自北海岸和巴斯海峽群島的黑唇鮑魚的性別、長度、重量等等。本實驗將分別使用 K-means 、階層式分群、 DBSCAN 進行群聚分析，比較各方法分群所花費時間，並使用 Purity 、 ACC 、輪廓係數（Silhouette Coefficient），分群品質衡量指標比較分群結果。

關鍵字：K-means；階層式分群；DBSCAN；群聚分析；機器學習

一、緒論

1.1 動機

世界上不管是在陸地上還是大海之中都充滿了無數的動植物，伴隨著地球好幾億個歲月，但大多數的人們並不是植物或生物學家，也並沒有足夠充裕的時間能夠認識廣大植物界、生物界的成員，本實驗前者將使用 Iris Data Set 這份完整的資料集做群聚分析，透過數據特徵值查看是否能將鳶尾的亞屬做良好的區分。後者使用 Abalone Data Set 資料集做群聚分析，透過鮑魚數據特徵查看是否可以將年齡屬性作良好區分。

1.2 目的

同屬或同種的動植物有時候僅僅是透過人眼觀察和專業經驗判斷時，仍有失誤的可能性，此時，配合資料集內的特徵觀察研究數據結果，可以讓人們更為容易地判別種類和年紀的差異，本實驗前者使用 Iris Data Set ，後者使用 Abalone Data Set 做出良好的群聚分析，藉此能夠讓不管是否擁有專業的學術知識，卻想以特徵數據判別動植物種類的人們，可以輕易辨別及認識。

二、資料集

2.1 真實資料集

本實驗的資料來源於 UCI Datasets 的 Iris Data Set 資料集以及 Abalone Date set 資料集。Iris Data Set 資料集是 Barry Becker 從 1994 年的人口普查資料庫中提取出來的。本實驗將前者的 Iris.data 的資料取出，共有 150 筆資料，資料集有 5 個屬性欄位，其中 4 個特徵屬性值都是實數，分別記錄鳶尾花瓣及花萼的長寬數據，最後一欄為類別標籤，屬於鳶尾屬下的三個亞屬，分別為山鳶尾（setosa）、變色鳶尾（verscicolor）和維吉尼亞鳶尾（virginica）。由於筆數過多無法全部呈現，部分資料內容如表 1 所示；後者 Abalone 資料集是由塔斯馬尼亞鮑魚的種群生物學提取出來。共有 4177 筆資料，資料集中有 9 個屬性欄位，其中 7 個欄位為連續屬性。分別紀錄鮑魚的重量、長度

、直徑等數據，另外 2 欄為離散型資料，分別紀錄鮑魚的年齡與性別，由於筆數過多無法全部呈現，部分資料內容如表 2 所示

表 1

Iris Data Set 資料集

Sepal_length	Sepal_width	Petal_length	Petal_width	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
…	…	…	…	…

表2

Abalone Data Set 資料集

sex	Length	Diameter	Height	Whole weight	…
M	0.455	0.365	0.095	0.5140	…
F	0.350	0.265	0.090	0.2255	…
…	…	…	…	…	…

三、方法

3.1 實作說明

本實驗針對 Iris Data Set 的處理分為了四個階段。第一階段做前置處理，將類別標籤的文字轉換成數字以利對結果做比較。第二階段畫出兩種視覺圖，一種是依據花瓣的長寬，另一種則是依據花萼的長寬。第三階段分別建構出 K-means、階層式分群以及 DBSCAN 三種模型，分為三群，分別計算各模型分群所花費之時間，並採用 Purity 衡量分群品質的指標。最後畫出階層式分群的階層樹（Dendrogram）。

Abalone Data Set 的處理分為四個階段，第二階段做前置處理，將性別欄位轉換成數字以利於後續做實驗，並將Rings設為標籤欄位，並將除標籤欄位的值都標準化，降低因資料尺度而造成的問題。第三階段分別建構出 K-means、階層式分群以及 DBSCAN 三

種模型，各自分為三群。分別計算個模型分群所花費之時間，並採用輪廓係數 (Silhouette Coefficiency) 與 accuracy 衡量分群品質的指標。最後畫出階層分群數階層分階樹。

3.2 操作說明

首先在 UCI Dataset 網站上找到 Iris Data Set 資料集，下載 Iris.data 檔案引入至實驗環境 jupyter notebook，將類別標籤透過 factorize()[0] 的方式把文字轉換為數字型態，以利與分群結果做比較。引入 matplotlib.pyplot、seaborn，利用花瓣長寬的數據組和花萼長寬的數據組分別畫出兩組視覺圖，接著建構 K-means、階層式分群、DBSCAN 三種模型，透過 time 工具計算不同模型的分群花費時間，並分別畫出分群結果的視覺圖以及印出分群結果利於與原先結果做比較，再運用 Purity 指標衡量分群結果。最後利用 scipy.cluster.hierarchy 工具畫出階層樹（Dendrogram）。

Abalone Data Set 同樣從 UCI Dataset 網站進行下載，將下載的 Abalone.data 檔案引入至實驗環境 jupyter notebook，將 Rings 欄位設定成標籤欄位，並將該欄位分成三群，再透過 actorize()[0] 的方式把文字轉換為數字型態，以利於與分群結果比較。並將性別欄為轉成數值，再透過標準化解決尺度造成的問題，接著建構 K-means、階層式分群、DBSCAN 三種模型，透過 time 工具計算不同模型的分群花費時間，並印出分群結果，再運用輪廓係數 (Silhouette Coefficiency) 與 accuracy 指標衡量分群結果。最後利用 scipy.cluster.hierarchy 工具畫出階層樹（Dendrogram）。

四、實驗

4.1 前置處理

Iris Data Set 資料集本身完整並無缺失值，且特徵屬性皆為實數，針對類別標籤採用了 factorize()[0] 的方法將文字轉換成數字，利於後續做群聚分析後的結果比較。

Abalone Data Set 資料集本身在原始資料中已經刪除缺失值，將年紀以分佈圖，如圖 1 所示，並將其分成三群，再將性別 One - Hot Encoding 轉成數值型態，最後使用 Min-Max。

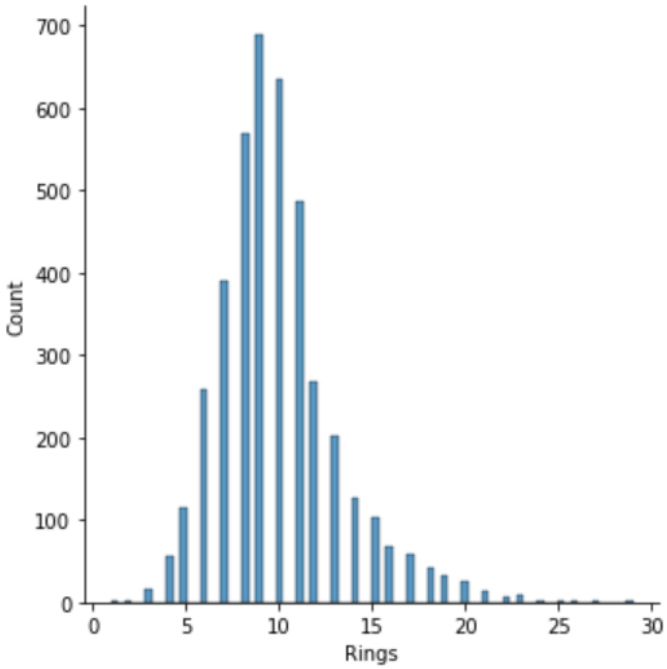

圖 1 Rings欄位年齡分布圖

4.2 實驗設計

在 Iris Data Set 中，建構出三個模型，分別為 K-means、階層式分群、 DBSCAN 模型，並分為三群，分別計算出不同模型下分群所花費之時間，以及透過 Purity 指標衡量分群品質。其中，K-means 和階層式分群的參數 n_clusters 皆設為 3，DBSCAN 的參數在預設的 eps = 0.5、min_samples = 5 時即能將資料集的內容分為三群。

在 Abalone Data Set 中，同樣建構三個模型，分別為 K-means、階層式分群、DBSCAN 模型。K-mean採用計算各個 n_clusters 參數的 SSE 值，如圖 2 所示。結果顯示當 n_clusters 為 3 時 SSE 的下降斜率會從傾斜大幅轉為平緩，固選擇 3 為 n_clusters 參數值。階層式分群下分別去計各算各 n_clusters 參數值的輪廓係數，如圖 3 所示。結果顯示在 n_clusters 為 3 時會得到最好的績效。DBSCAN採預設值 eps = 0.5、min_samples = 5 進行分群。

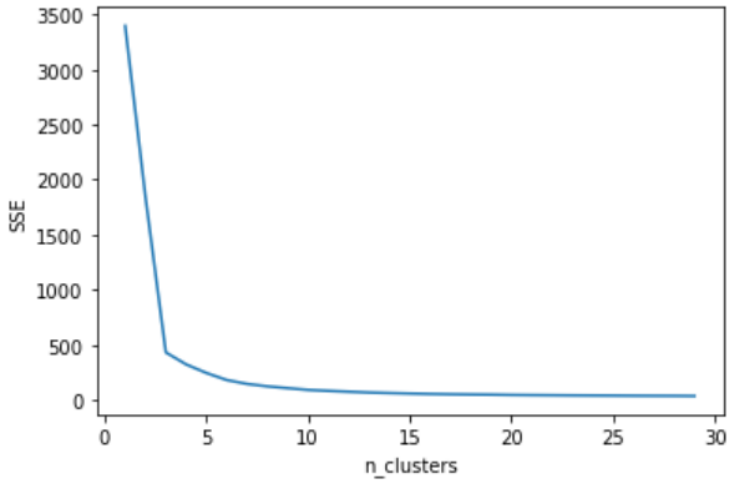

圖 2 K-mean各 n_clusters 下 SSE 值折線圖

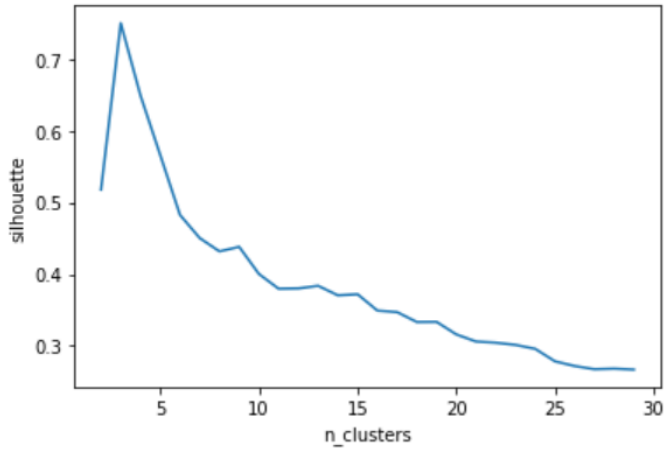

圖 3 階層式分群各 n_clusters 下 輪廓係數折線圖

4.3 實驗結果

Iris Data Set 的 K-means、階層式分群、DBSCAN 的實驗結果根據花瓣長寬呈現之視覺圖分別如圖 4、圖 5、圖 6 所示。分群執行時間如表 3 所示。Purity 衡量指標結果如表 4 所示。圖 7 則為階層式分群之階層樹（Dendrogram）。

Abalone Data Set 的 K-means、階層式分群、DBSCAN 的實驗結果根據鮑魚的直徑與重量呈現視覺圖，分別為圖 8、圖 9、圖 10。執行，分群所需時間如表 5 所示。使用輪廓係數與 accuracy 衡量指標的結果如表 6 所示。圖 11 為階層式分群之階層樹。

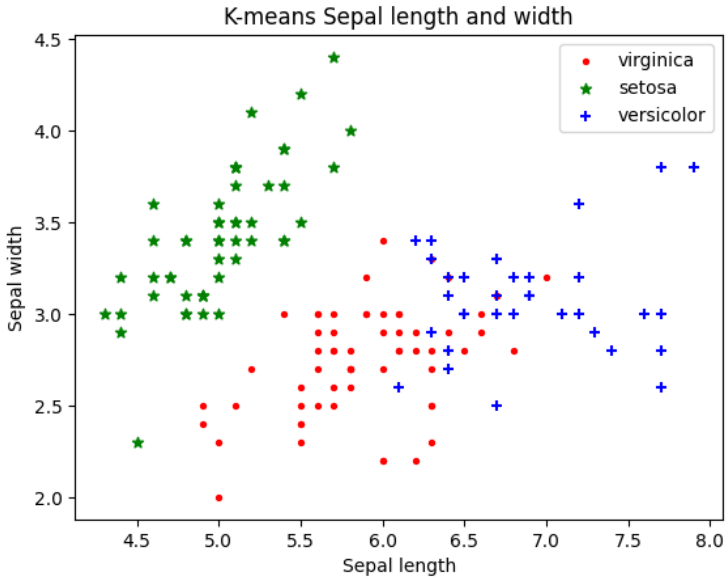

圖 4 Iris Data Set K-means 分群結果之花瓣長寬視覺圖

將 K-means 的 n_clusters 設為 3 時，由圖 4 可知 K-means 對此資料集有不錯的分群結果。

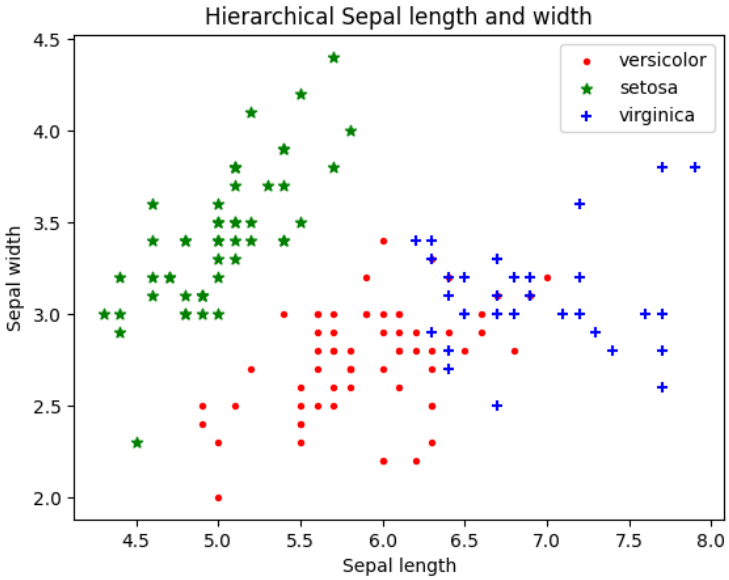

圖 5 Iris Data Set 階層式分群結果之花瓣長寬視覺圖

將階層式分群的 n_clusters 設為 3時，由圖 5 可知階層式分群對此資料集也有不錯的分群結果。

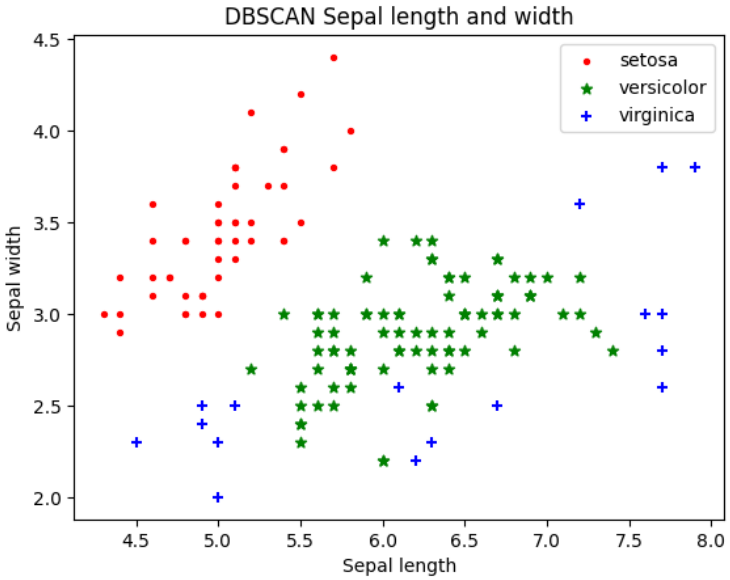

圖 6 Iris Data Set DBSCAN 分群結果之花瓣長寬視覺圖

由圖 6 可知 DBSCAN 對此資料集較無法正確分群，此模型將 virginica 視為了雜訊。

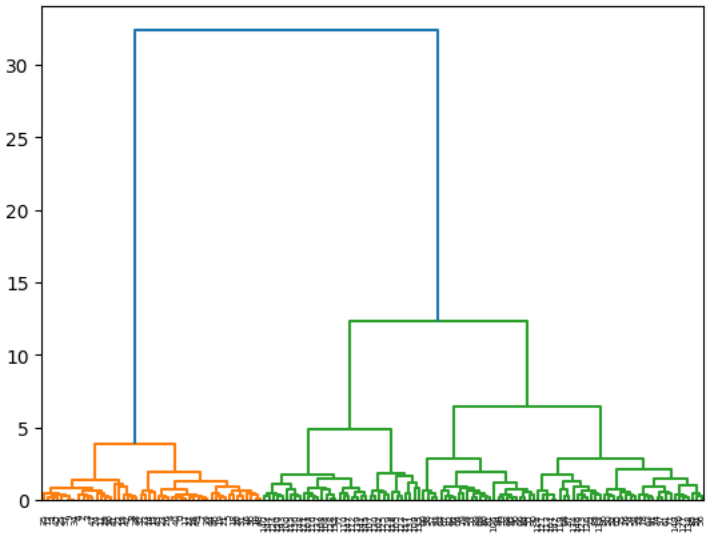

圖 7 Iris Data Set 階層樹

表 3

Iris Data Set DBSCAN 使用 K-means、階層式分群、DBSCAN 分群之分別執行時間

	K-means	階層式分群	DBSCAN
分群花費時間(s)	.070	.002	.003

由表 3 可知 K-means 是三種模型之中分群所需花費時間最長的模型。

表 4

K-means、階層式分群、DBSCAN 各模型之 Purity 純度

	K-means	階層式分群	DBSCAN
Purity	.893	.893	.687

由表 4 可知 Iris Data Set 資料集在 DBSCAN 模型下的分群結果並不優秀。

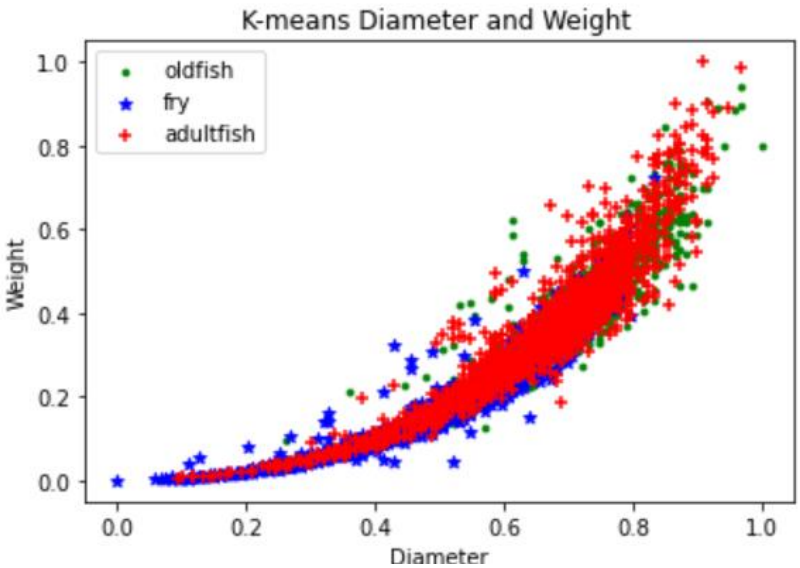

圖 8 Abalone Data Set K-means 直徑與重量分群結果視覺圖

將 K-means 的 n_clusters 設為 3 時，由圖 8 可知分群在 adultfish 與 old fish 群與真實結果有所差異。

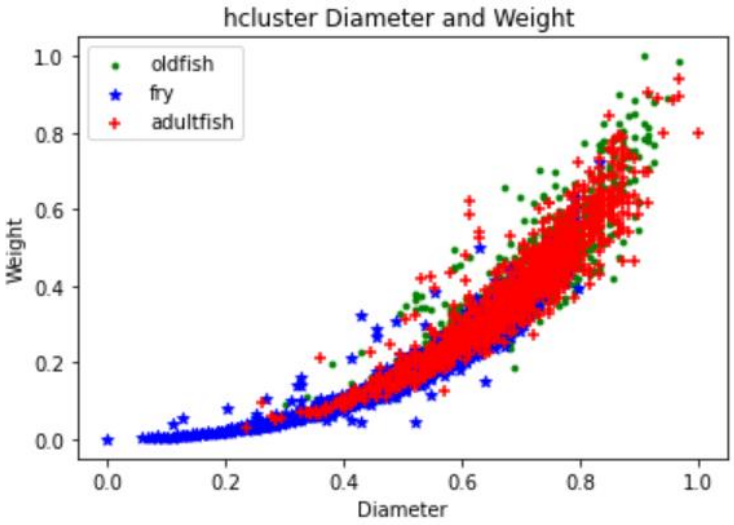

圖 9 Abalone Data Set 階層式分群直徑與重量分群結果視覺圖

將階層式分群的 n_clusters 設為 3時，由圖 9 可知階層式分群同樣在 old fish 群與真實結果有所差異。

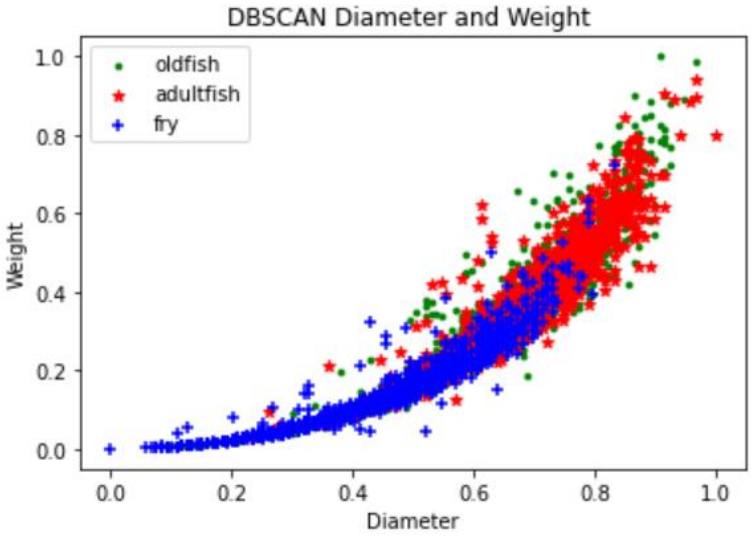

圖 10 Abalone Data Set DBSCAN直徑與重量分群結果視覺圖

由圖 10 得知，Abalone Datea Set 在 DBSCAN 上與真實資料有顯著的落差，故此資料集較不適用此演算法。

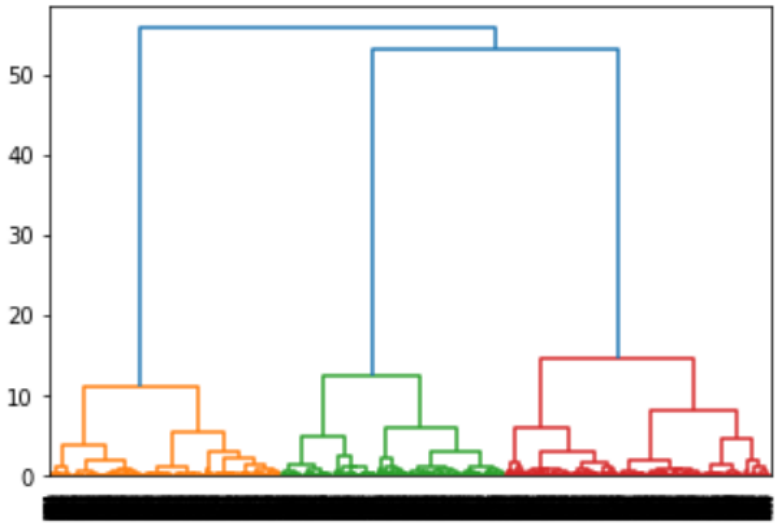

圖 11 Abalone Data Set 階層樹

由表 5 可以得知 Abalone Data Set 使用三種方法分群，依時間短到長分別是，k-meams、DBSCAN、階層式分群。

表 5

Abalone Data Set 使用 K-means、階層式分群、DBSCAN 分群之分別執行時間

	K-means	階層式分群	DBSCAN
分群花費時間(s)	.052	.615	.362

由表 6 可以得知 Abalone Data Set 使用三種方法分群，依 Silhouette係數可以得知，k-meams 跟階層式分群有相同的績效，較差的是 DBSCAN，而從 ACC 中可以得知，階層式分群與其他兩個方法有較好的績效。

表 6

K-means、階層式分群、DBSCAN 各模型之 輪廓係數、ACC

	K-means	階層式分群	DBSCAN
Silhouette係數	.750	.750	.710
accuracy_score	.487	.516	.329

五、結論

本實驗中，透過不同的模型將 Iris Data Set 分群，由實驗結果可以得知，K-means 的分群速度相較於階層式分群以及 DBSCAN 來說需要花費較多的時間，但是從視覺化圖像和 Purity 衡量分群品質指標中可以得知，K-means 和階層式分群皆有良好的績效，將上述的速度和正確性綜合審視，Iris Data Set 資料集在使用階層式分群模型上能夠得到最佳的分群結果。

在 Abalone Data Set 資料集中使用了不同的方法分群，由實驗結果可以得知，階層式分群花費的時間最多，K-means 花費時間最少，而績效從輪廓係數得知，K-meams 和階層式分群有較好的績效，但從 ACC 中得知，階層式分群與其他兩種方法有較好的績效。要分析 Abalone Data Set 資料集要有較快的時間可以選擇 K-meams，若要有較好的績效可以選擇使用階層式分群。

參考文獻

Joeyajames. (2021, October 29). *Python/Iris_Dataset.Ipynb at Master*. GitHub. https://github.com/joeyajames/Python/blob/master/Iris%20Dataset/Iris_Dataset.ipynb

Kuka. (2015, December 2). *Python Clustering "purity" Metric*. Stack Overflow. https://stackoverflow.com/questions/34047540/python-clustering-purity-metric

中文文獻

Tonykuoyj。（2016 年 12 月 24 日）。*[第 24 天] 機器學習（4）分群演算法*. IT邦幫忙. https://ithelp.ithome.com.tw/articles/10187314

浪客竹馬。（2018 年 8 月 22 日）。*鳶尾花三種聚類算法（K-Means,AGNES,DBScan）的python實現*。台部落。 https://www.twblogs.net/a/5b7c80fb2b71770a43db4af3

大數據界olu。（2022年 4 月 8 日）。*實踐 | K-Means聚類（使用鳶尾花數據集）*。Chinese Software Developer Network。 https://blog.csdn.net/jiangti_ng/article/details/123644380