資料探勘專題作業三

利用 Python 軟體實作群聚分析

指導教授:

許中川 教授

成員:

M11123026 林宥昇

M11123047 劉穎謙

M11123055 蕭旭朝

日期:

2022 年 12 月 13 號

摘要

本研究使用 Iris 資料集、Estimation of obesity levels based on eating habits and physical condition 資料集透過 K-means、階層式分群、DBSCAN，將資料分成 3 群並產生階層樹，並且比較分群所花費時間，Iris 資料集使用 Purity 指標衡量分群品質，另一個 Estimation of obesity levels based on eating habits and physical condition 資料集使用 Purity 指標及 Silhouette Coefficient 與 Calinski Harabasz Score 衡量分群品質。

關鍵字：鳶尾花、資料探勘、K-means、階層式分群、DBSCAN、肥胖

一、緒論

1.1 動機

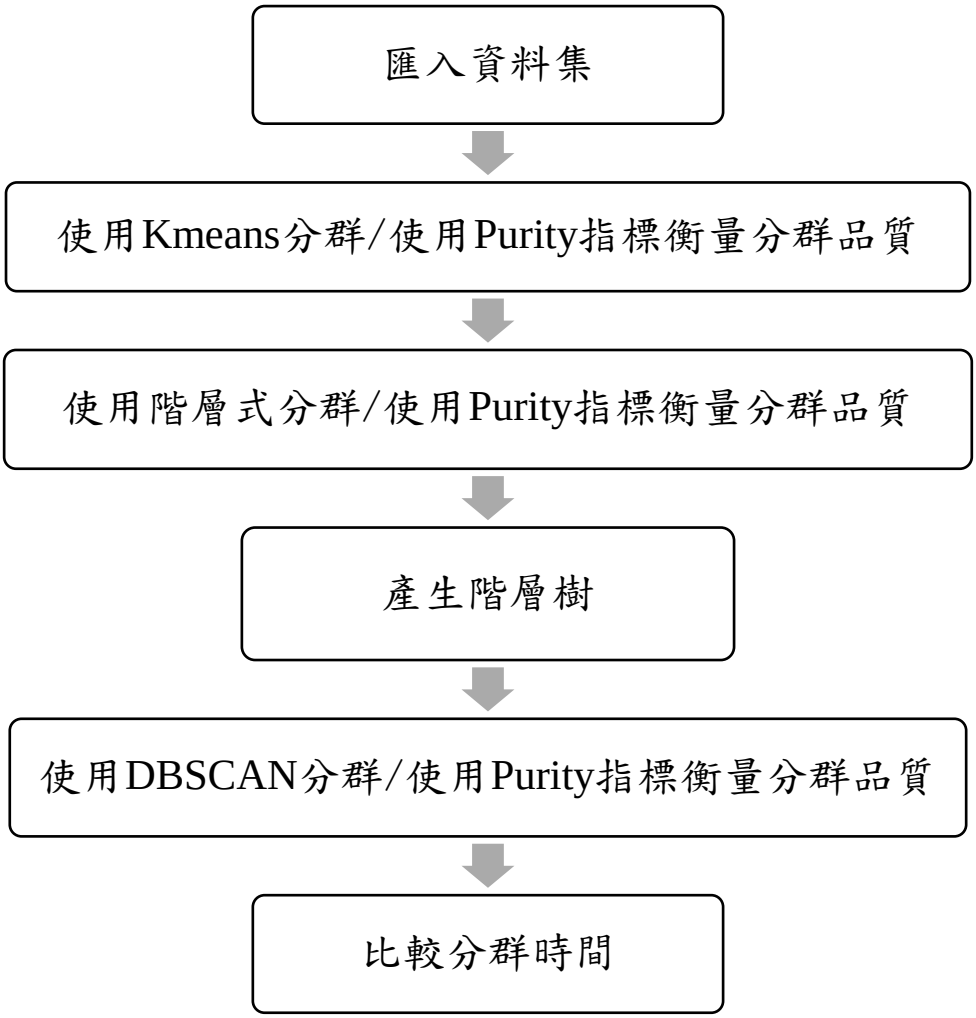
分群法是相對於分類法的另外一種資料探勘技術。分群法也是用來將資料做區分的，差別在於原本的資料都是未經過類別區分的。因為是未知類別的資料集進行區分所以也被稱為非監督式學習，分群法針對沒有預先定義好類別的資料分組，本研究使用 Iris 資料集和 Estimation of obesity levels based on eating habits and physical condition 資料集來探討 Purity 指標及 Silhouette Coefficient 與 Calinski Harabasz Score 衡量分群品質。

1.2 目的

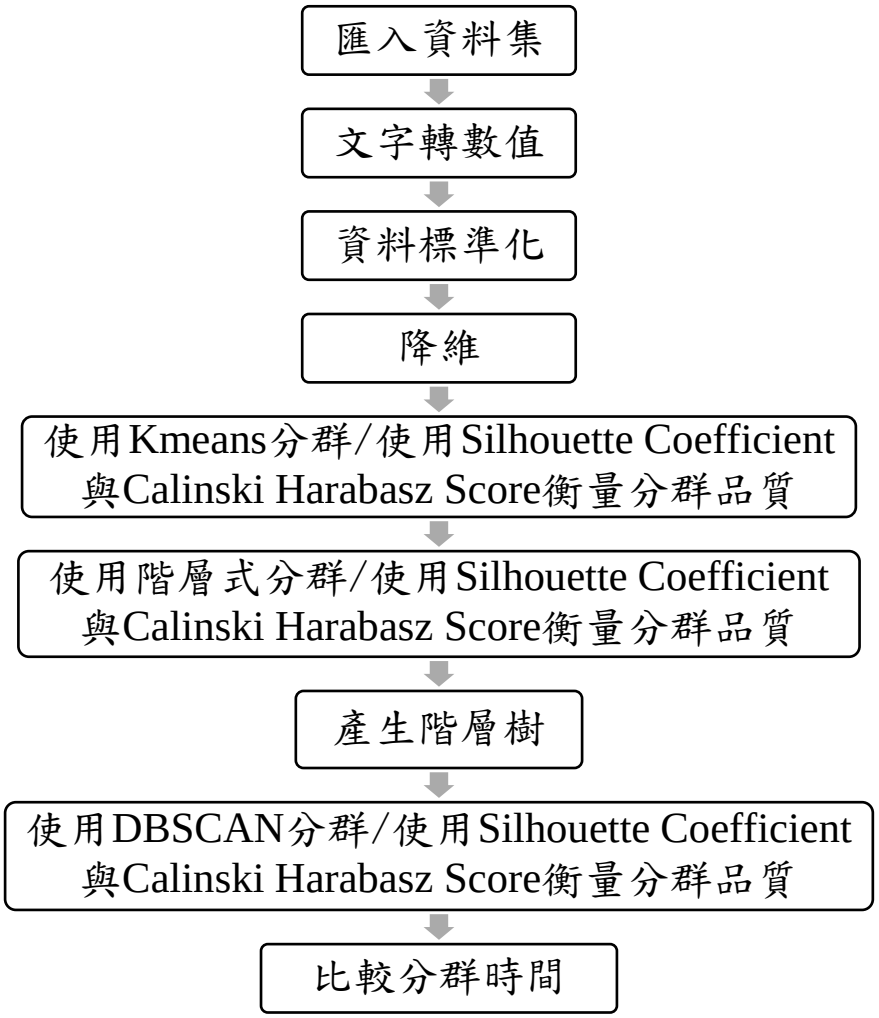
本次研究使用 Iris 資料集和 Estimation of obesity levels based on eating habits and physical condition 資料集，通過 K-means、階層式分群、DBSCAN 來比較分群所花費時間並且使用 Purity 指標及 Silhouette Coefficient 與 Calinski Harabasz Score 衡量分群品質。

二、方法

2.1 Iris 資料集



2.2 Estimation of obesity levels based on eating habits and physical condition 資料集



三、實驗

3.1 Iris 資料集

3.1.1 Iris 資料集

sepal length in cm	花萼長度（厘米）
sepal width in cm	花萼寬度（厘米）
petal length in cm	花瓣長度（厘米）
petal width in cm	花瓣寬度（厘米）
setosa	山鳶尾
versicolor	變色鳶尾
virginica	維吉尼亞鳶尾

3.1.2 Estimation of obesity levels based on eating habits and physical condition 資料集

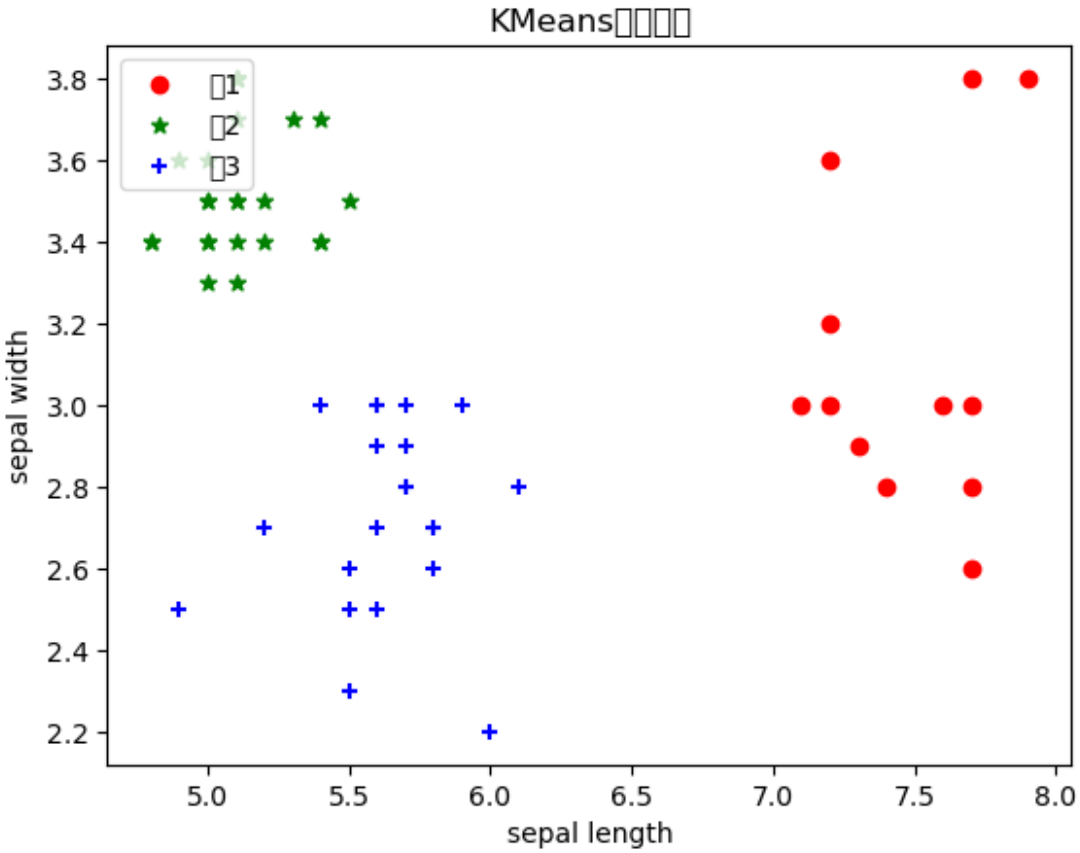
Gender	性別
Age	年齡
Height	身高
Weight	體重
Family history with overweight	家庭成員是否超重
FAVC	經常食用高熱量食物
FCVC	食用蔬菜的頻率
NCP	主餐次數
CAEC	兩餐之間的食物消耗
SMOKE	是否抽煙
CH2O	每日用水量
SCC	卡路里消耗監測
FAF	體力活動頻率
TUE	使用技術設備的時間
CALC	飲酒
MTRANS	使用的交通工具
NObeyesdad	肥胖程度

3.2 實驗設計

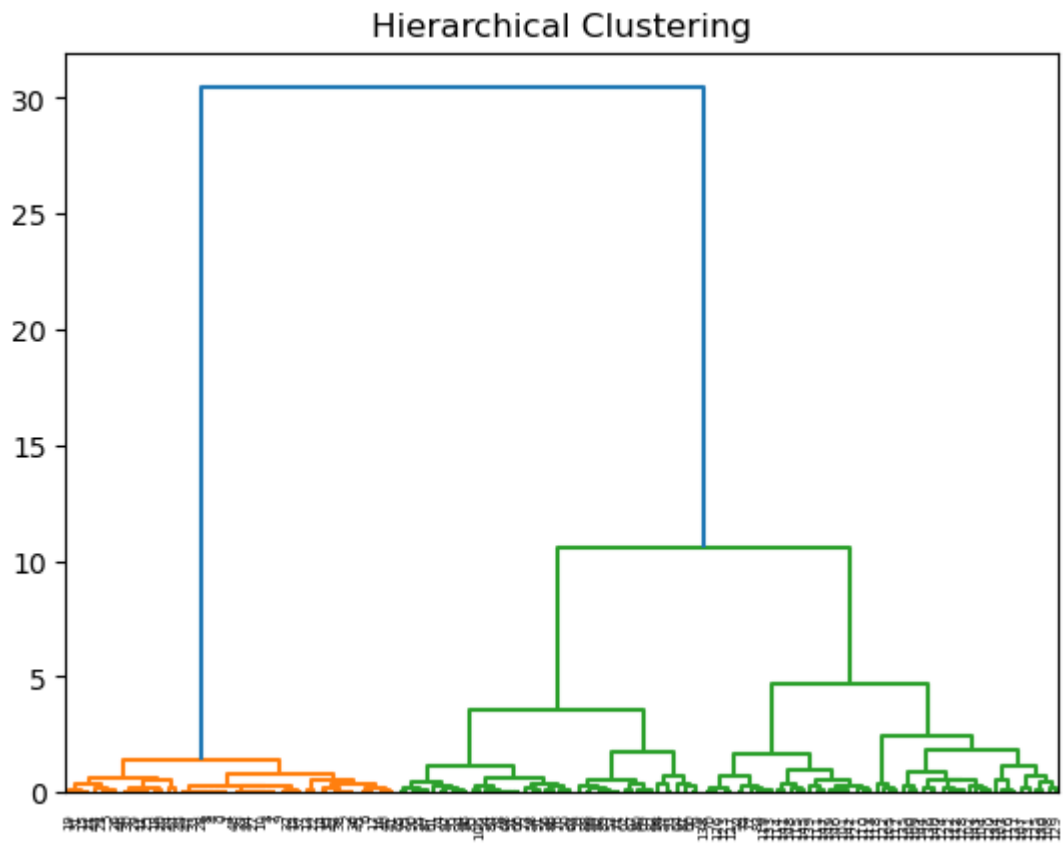
針對 Iris 資料集及 Estimation of obesity levels based on eating habits and physical condition 資料集分別使用 KMeans、階層式分群、DBSCAN 三種分類方式進行分群，計算 Purity 指標及 Silhouette Coefficient 與 Calinski Harabasz Score 衡量分群品質及產生階層樹，最後比較三種分群方式的時間。

3.3 實驗結果

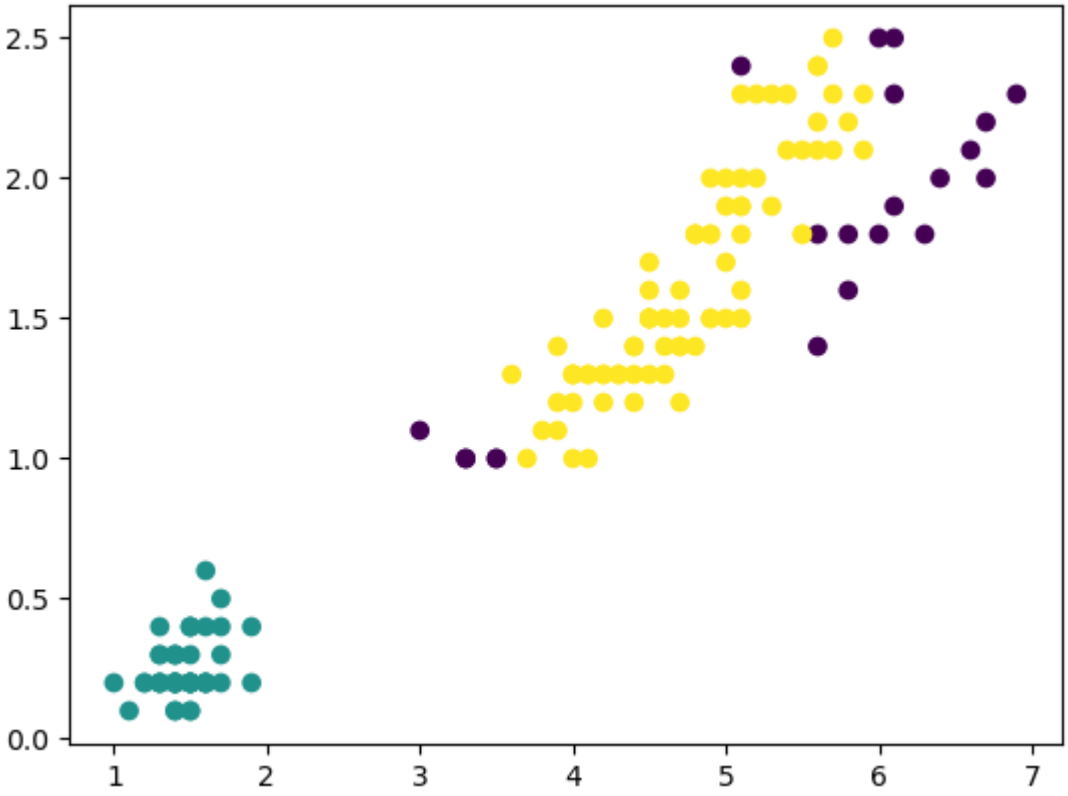
3.3.1 Iris 資料集



KMeans	Purity	花費時間
	0.973	0.2 秒

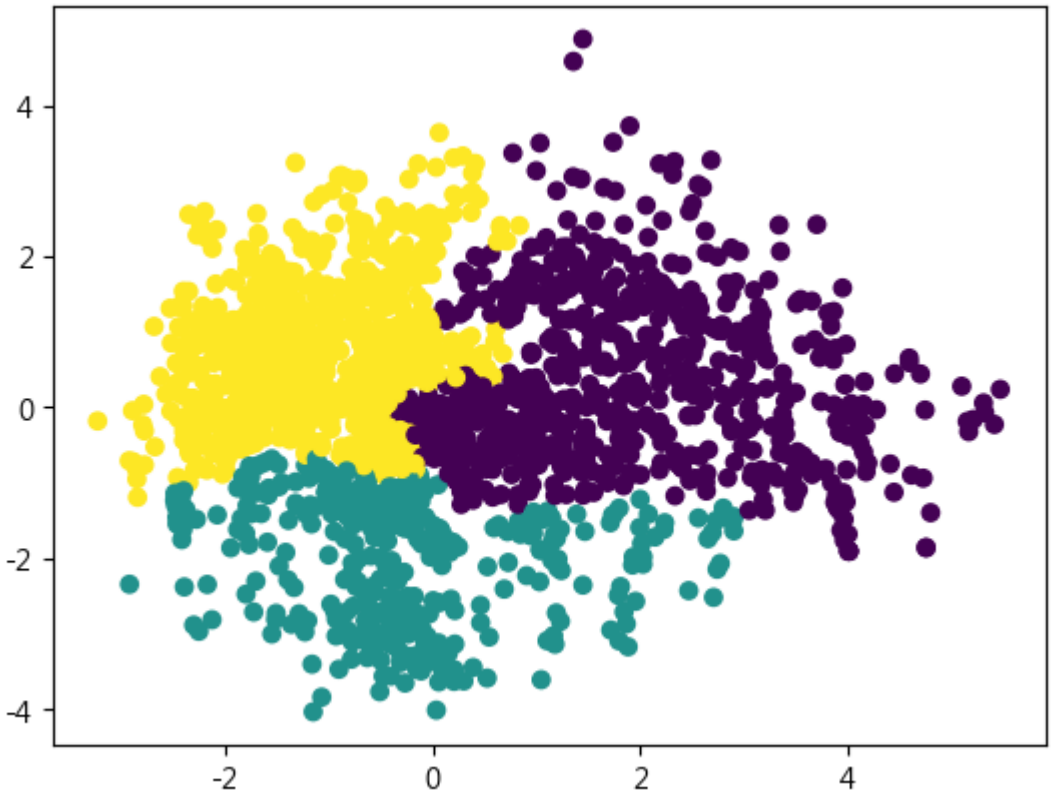


Hierarchical	Purity	花費時間
	0.96	1.8 秒

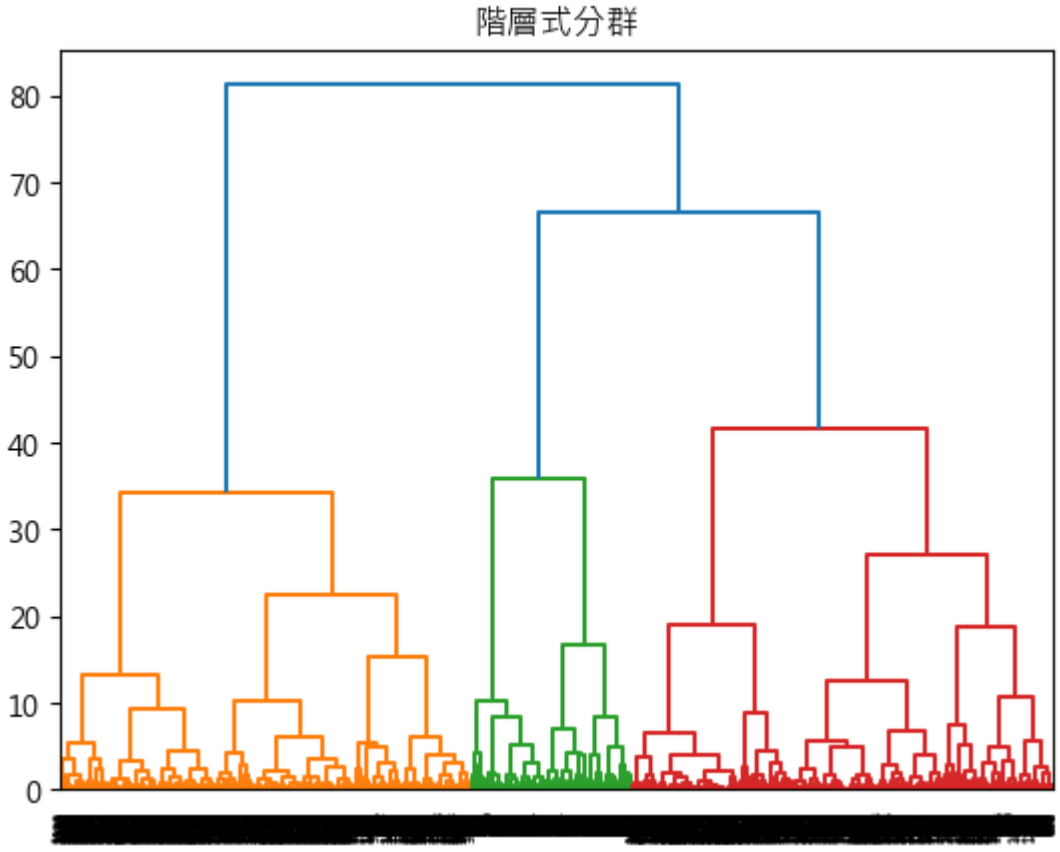


DBSCAN	Purity	花費時間
	0.74	0.2 秒

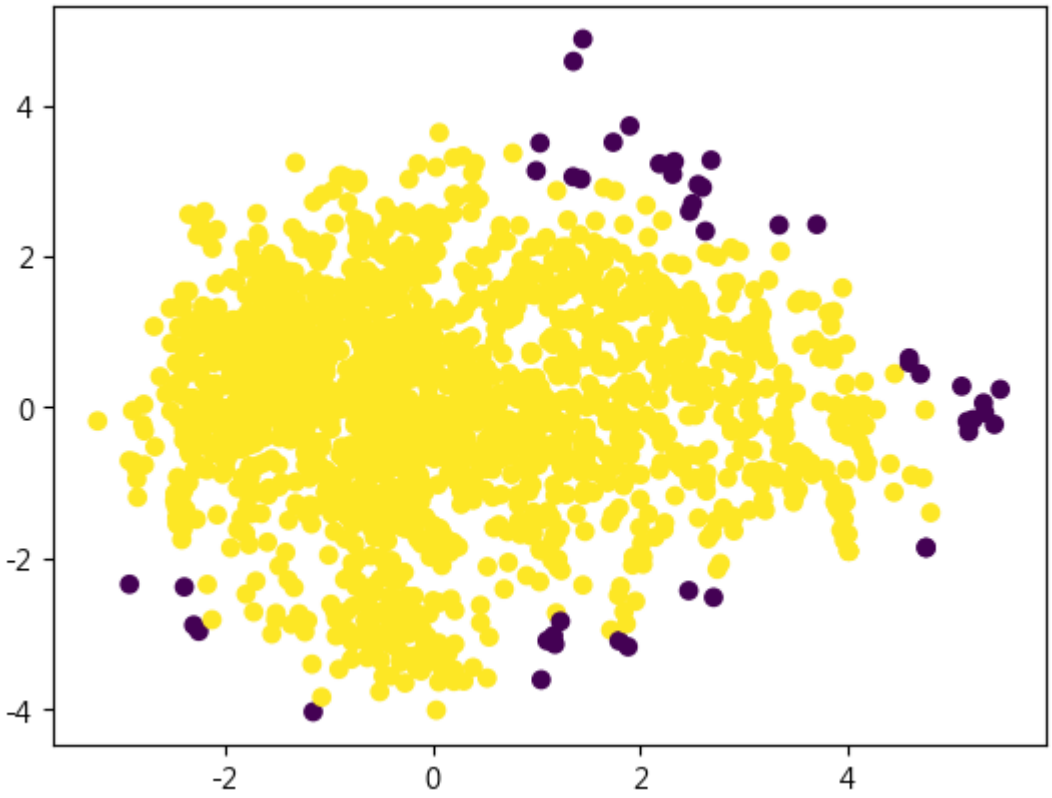
3.3.2 Estimation of obesity levels based on eating habits and physical condition 資料集



KMeans	Silhouette Coefficient	Calinski Harabasz Score	花費時間
	0.4024	1787.410	0.105 秒



Hierarchical	Silhouette Coefficient	Calinski Harabasz Score	花費時間
	0.341	1312.942	0.14 秒



DBSCAN	Silhouette Coefficient	Calinski Harabasz Score	花費時間
	0.384	59.566	0.0193 秒

四、結論

4.1 Iris 資料集

在此資料集中，使用 KMeans 分群，純度明顯優於其他兩種分群方式，階層式分群次之，DBSCAN 最低，DBSCAN 與 KMeans 執行時間相同，階層式分群次之，綜合比較之下 KMeans 最適於此資料集。

4.2 Estimation of obesity levels based on eating habits and physical condition 資料集

在此資料集中，使用 KMeans 分群，品質明顯優於其他兩種分群方式，階層式分群次之，DBSCAN 最低，但 DBSCAN 執行時間是三個分群方式之間最短，綜合比較之下 KMeans 最適於此資料集。

五、參考文獻

10 程式中(2021 年)。[Day 7] 非監督式學習-降維。

https://ithelp.ithome.com.tw/m/articles/10267685

10 程式中(2021 年)。[Day 3] 你真了解資料嗎？試試看視覺化分析吧！

https://ithelp.ithome.com.tw/m/articles/10264416

PyInvest(2020 年 7 月 17 日)。[Python 實作] 密度聚類 DBSCAN。

https://pyecontech.com/2020/07/17/python_dbscan/

10 程式中(2019 年 9 月 18 日)。[Day 6] 非監督式學習 K-means 分群。

https://ithelp.ithome.com.tw/articles/10266672

LoveMIss-Y(2019 年 7 月 3 日)。基於 sklearn 的聚類算法的聚類效果指標

https://www.twblogs.net/a/5d1bbc22bd9eee1ede05c0d2