# 透過分類與分群分析含有未知標籤的醫學資料集
# gene expression cancer RNA-Seq Data Set

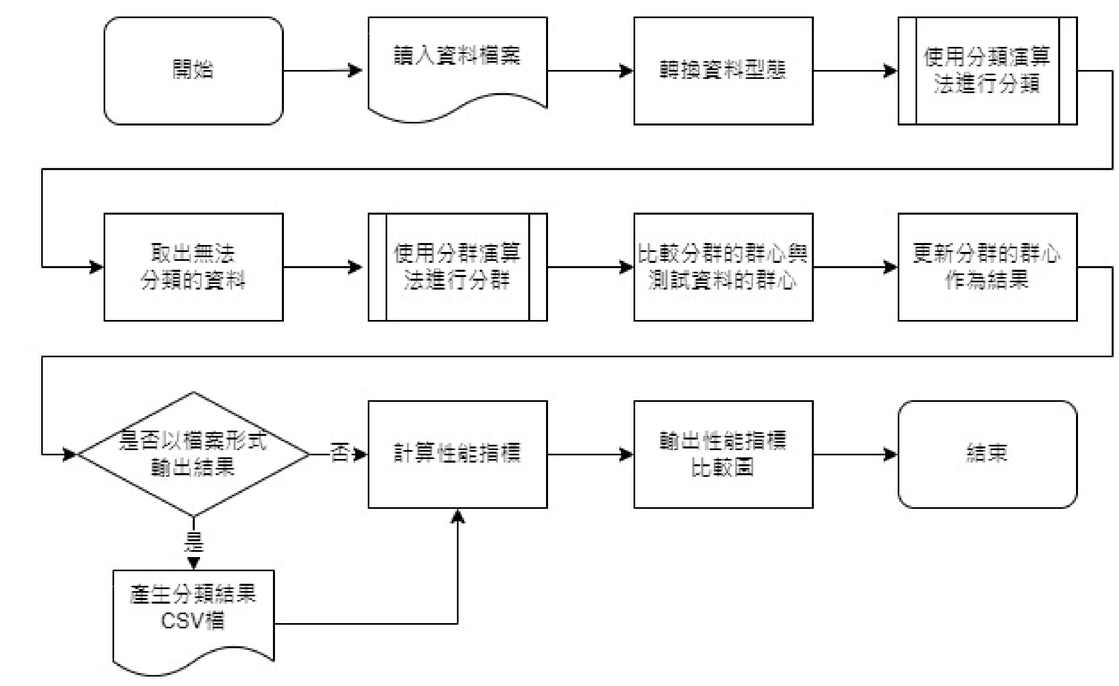
指導老師：蔡崇煒 教授
團隊成員：鄭璟翰、郭晏涵

## 一、前言

隨著大數據的時代來臨，各種感應裝置普及，資料的蒐集已經也比以往還要方便許多；然而資料快速增加導致了資料的數量與維度都大幅上升，而醫學上的數據也是如此。而過往的分類往往需要先透過訓練資料所產生的分類器，在面對訓練資料中所沒有的標籤時，並沒有辦法正確分類。因此這次期末報告就採用「先分類後分群」的方法來進行：先依照訓練資料訓練出分類器，對於分類器不確定的資料先分類到未知類別，蒐集所有未知類別的資料後再進行分群，當作為分類結果。

這次期末報告我們所選擇的資料集是Gene Expression Cancer RNA-Seq Dataset，每一筆資料總共有大約兩萬筆有關基因的資料；訓練資料中有三種分類結果，而測試資料當中多加入了兩種分類結果。分析的方法則是透過python實作兩種分類器K-Nearest Neighbor、Random Forest(透過scikit-learn完成)以及兩種分群方法K-Means、DBSCAN來分析資料集，因此總共會產生六種分類結果，包括兩種分類器加上兩種分類器乘上兩種分群方法。

## 二、程式流程圖




## 三、分類器選擇

為了要適應測試資料集中有未知分類的結果，需要選擇有客觀指標分類器，才能有效的評斷一筆資料是否分類無效，進而產生出未知標籤的資料。這裡選擇KNN以及Random Forest作為分類器：

(1)KNN分類器是藉由找出距離最小K筆資料的方法依照距離分類，因此可以藉由直接判斷和其他資料的距離來判斷是否為無法分類的資料。藉由計算訓練資料中所有點的距離取平均得知任兩點的平均距離約為200，因此我們設定將最小K筆資料距離總和若大於200 * K + 100，則判斷為未知的分類項目，若當K=9，則為前9筆的距離總和大於1900則設定為未知資料。

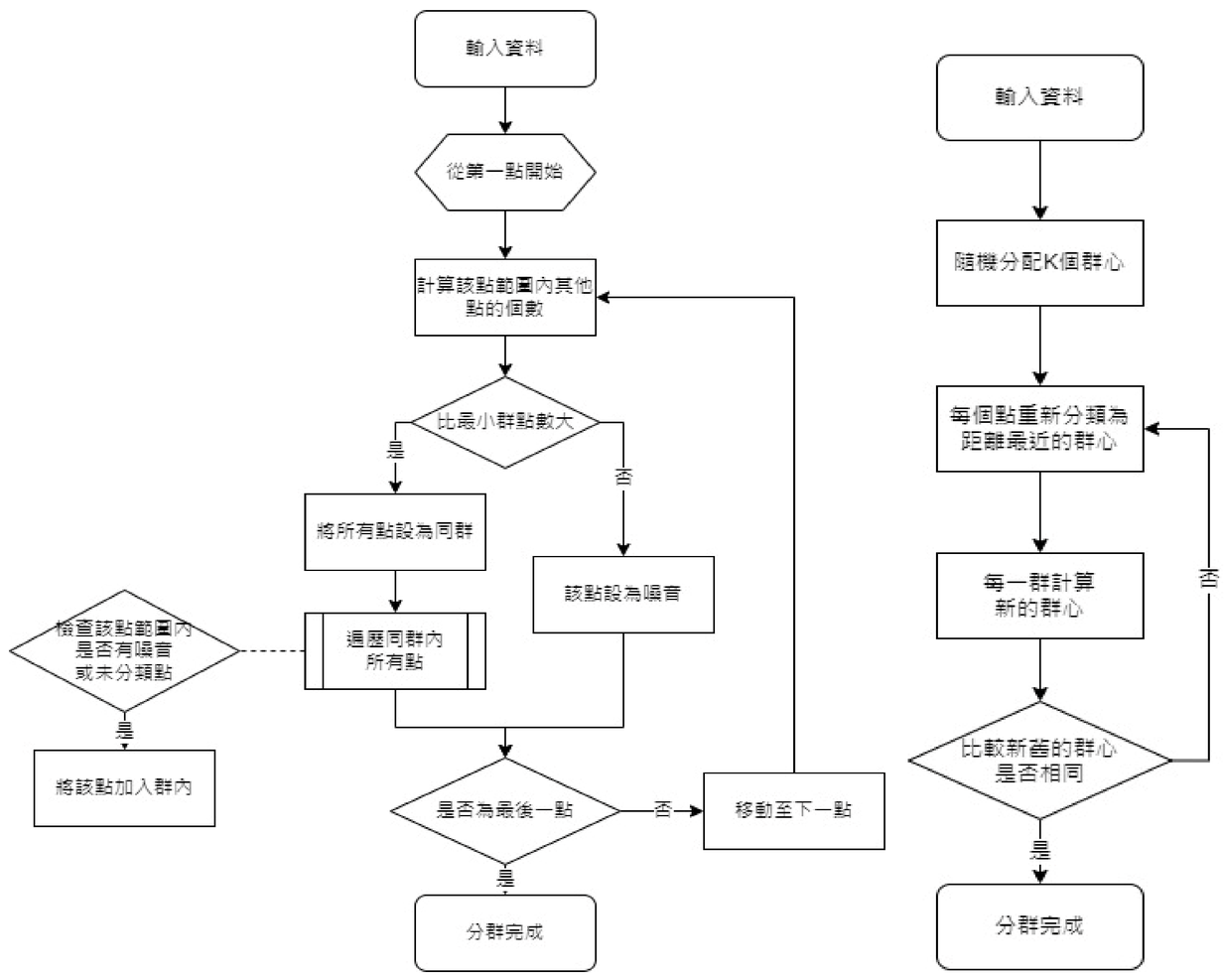
(2)Random Forest分類器是從訓練資料中選取子資料讓Decision Tree來訓練，最後由各個Decision Tree來進行投票。因為有投票的過程，各筆資料的分類結果就可以以機率的方式展現，因此就可以將機率過低的結果視為未知分類。藉由scikit-learn 實現的話則可以依照predict_pro所回傳的機率表來進行分類。若該筆資料之最高機率的分類項目沒有超過0.7，則將該筆資料設定為未知的分類項目。

## 四、分群

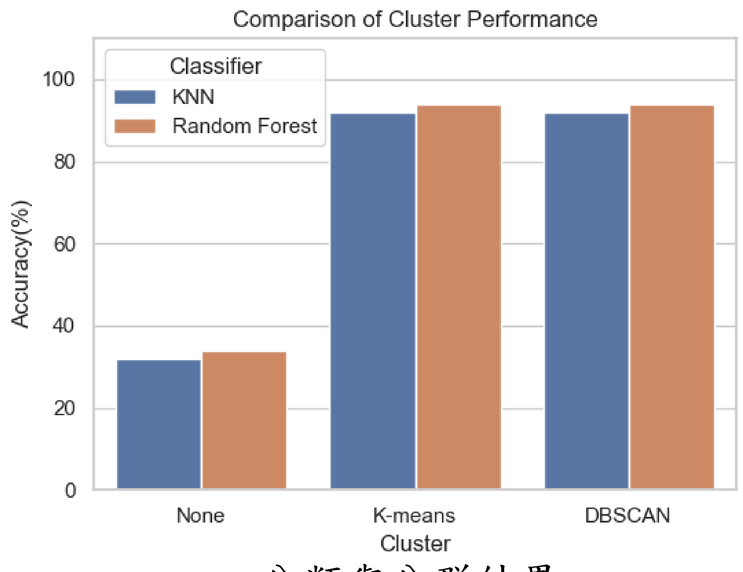
進行完上一階段的分類訓練後，接著對未知的項目進行分群的演算法，採用K-Means以及DBSCAN兩種分群演算法。

K-Means是一種將資料依照平均距離分割成K類別分群方法，其中K是hyper parameter。一般情況下，很難得知K要設定為多少才能正確分類，但在這次的測試的資料集中，我們可以得知未知的分類標籤總共有兩個，因此可以設定K=2。

DBSCAN，全名為Density-based spatial clustering of applications with noise，是一種基於密度來將資料進行分群的演算法。給定esp(高密度資料的範圍)以及minPts(組成高密度區的最小點數量)後，DBSCAN能把周圍的點組合成同一群，並標記出位於低密度區域的局外點(稱為噪音)。




DBSCAN流程圖　　K-Means流程圖




分類與分群結果

## 五、結果分析與結論

透過上圖分類與分群結果可以發現：

(1)、在只有進行classifier的情況下，分類器的準確度都只有1/3，因此可以推測約六成的測試資料都是訓練資料中所沒有的分類。

(2)、六種結果的準確度以random forest加上K-Means的準確率最高；但以時間複雜度而言，KNN所要耗費的時間太多，所以最有效率的方法為random forest加上clustering。

(3)、比較兩種分群K-Means的準確率稍微高一些，但我們認為是因為DBSCAN會將距離較遠的資料判定為噪音，才會導致準確率稍微降低。

| K-Means | DBSCAN |
|---|---|
| 依照到平均資料的距離分類 | 依照周圍資料的密度分類 |
| 沒有噪音所有點有結果 | 有噪音，周圍資料密度低時分為噪音 |
| 適合極端值少每個分類明確<br>不適合non-convex | 可適應噪音多的資料集 |
| 需事先設定分成幾類 | 不用先設定分成幾類 |

K-Means與DBSCAN比較