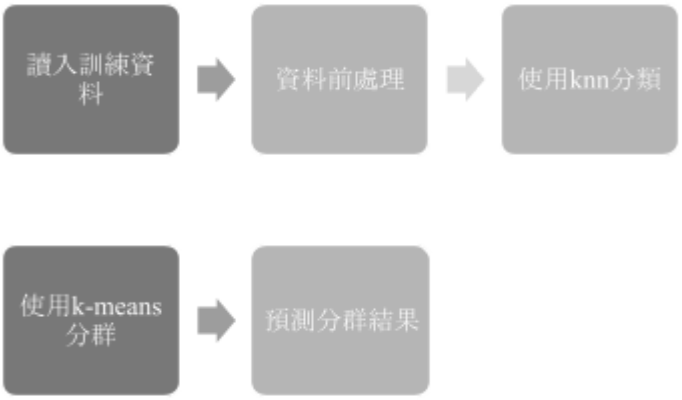
- ## 摘要

現代人口老化及不健康生活型態影響，每年罹癌人數不斷上升。由於早期病徵不明顯，使得許多患者錯過最佳的治療時期，更是提高了癌症的死亡率，長期佔據台灣十大死因之首的位置。這次實作我們利用RNA序列與癌症的關聯，以分類與分群的方法分析gene expression cancer RNA-Seq data set訓練模型，藉以分別各種癌細胞的特徵，並用於幫助療法的研發。訓練過程中使用KNN演算法分類已知資料，用K-Means演算法處理未知資料，增進訓練模型的靈活度與效率。

實作**KNN**流程圖

- ## 程式設計方式



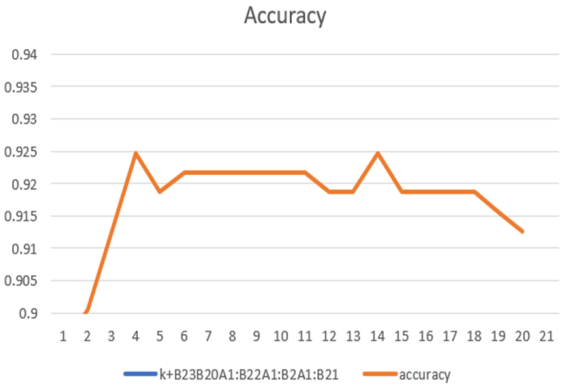
- ## 結果與分析

A. 執行結果分析

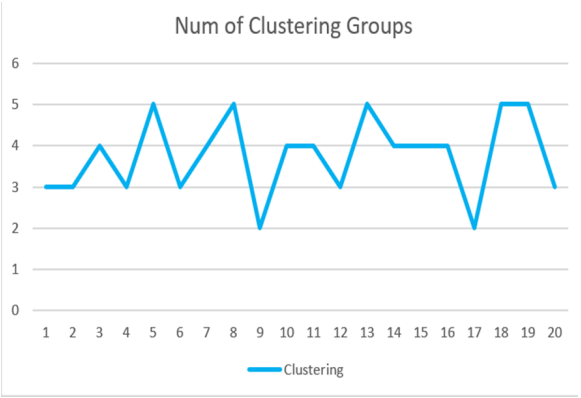
在使用KNN和K-means進行預測後，我們可以觀察一下整體的準確率，首先看一下預測的結果：

k	accuracy	分群次數
1	0.8945783132530121	3
2	0.9006024096385542	3
3	0.9126506024096386	4
4	0.9246987951807228	3
5	0.9186746987951807	5
6	0.9216867469879518	3
7	0.9216867469879518	4
8	0.9216867469879518	5
9	0.9216867469879518	2
10	0.9216867469879518	4
11	0.9216867469879518	4
12	0.9186746987951807	3
13	0.9186746987951807	5
14	0.9246987951807228	4
15	0.9186746987951807	4
16	0.9186746987951807	4
17	0.9186746987951807	2
18	0.9186746987951807	5
19	0.9156626506024096	5
20	0.9126506024096386	3
average	0.9173192771084338	3.75

※註：k值為knn的取點個數



由準確率之折線圖可以看出準確率大約都落在0.91~0.93之間，，另外經由觀察可以發現在k=1和k=2的時候準確率稍微偏低，推測應該是因為k值太小導致判斷分類的條件不夠充足，也可以發現在k=20左右準確率也有下降的趨勢，推測是因為測試資料中每個類別的資料總數可能不同，因此k值太大時會使預測結果被較多資料的類別影響，但因為此程式使用距離反比來加權，所以雖然會有影響但變動幅度應該不大。



上圖可以看出分群變動次數在2~5之間，而此程式設定之停止條件為50次，所以可以得知這個K-means演算法很有效率，而重新分群次數不多的原因可能是此程式在初始尋找形心時設定為找到距離適中的形心，因此可以很快區分出兩個類別，可是這種方法在分成三群以上時不一定適用，而且這種方法雖然能降低分群失敗的機率，但若形心選到outlier還是會造成分群的結果錯誤。