GROUP 8

資料探勘期末報告

B074011005 王濬瑋
B094020017 郭楨君
B094020046 黃奕瑋
B124020018 劉佳瑜
B124020027 陳闈霆

AGENDA

01 資料集介紹

02 程式流程

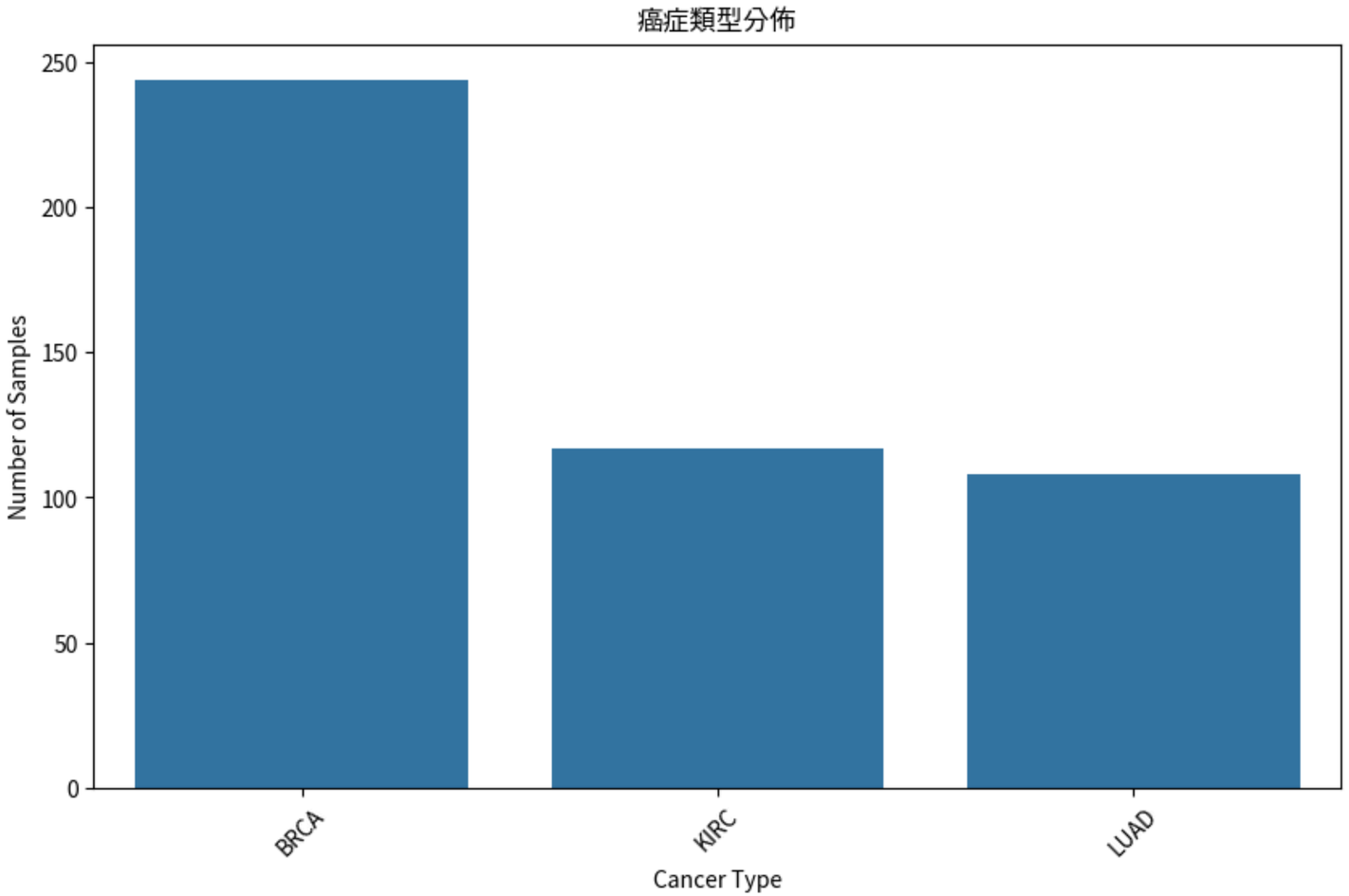
03 提升準確率

04 結論

資料集介紹

gene expression cancer RNA-Seq Data Set

[469 rows x 20532 columns]



程式流程

資料前處理

Random Forest
分類

Kmeans
分群

計算準確率

程式流程
STEP 1

資料前處理:

01

檢查缺失值

檢查列中是否有 NaN 值

02

檢查異常值

以四分位距來檢測異常值，
資料大於 Q3+1.5IQR 或
小於Q1-1.5IQR，視為離群值

03

處理異常值

用該欄位的平均值替換
異常值

程式流程
STEP 2

Random Forest 分類:

train_label: BRCA KIRC LUAD

程式流程
STEP 2

Random Forest 分類:

train_label: BRCA KIRC LUAD + unknown

程式流程
STEP 2

Random Forest 分類:

train_label: BRCA KIRC LUAD + unknown

data1 [0.1, 0.1, 0.8]

data2 [0.2, 0.3, 0.5]

data3 [0.1, 0.8, 0.1]

程式流程
STEP 2

Random Forest 分類:

train_label: BRCA KIRC LUAD + unknown

data1 [0.1, 0.1, 0.8] LUAD

data2 [0.2, 0.3, 0.5] LUAD

data3 [0.1, 0.8, 0.1] KIRC

程式流程
STEP 2

Random Forest 分類:

train_label: BRCA KIRC LUAD + unknown

if threshold = 0.7

data1 [0.1, 0.1, 0.8] LUAD

data2 [0.2, 0.3, 0.5] unknown

data3 [0.1, 0.8, 0.1] KIRC

max prob 小於 threshold 判斷為 unknown

程式流程
STEP 3

選擇 K 值

選擇K個
初始群心

計算資料到每個
群心的歐式距離

確認群心
不再變動

計算新的群心

將資料分配至
距離最近的群心

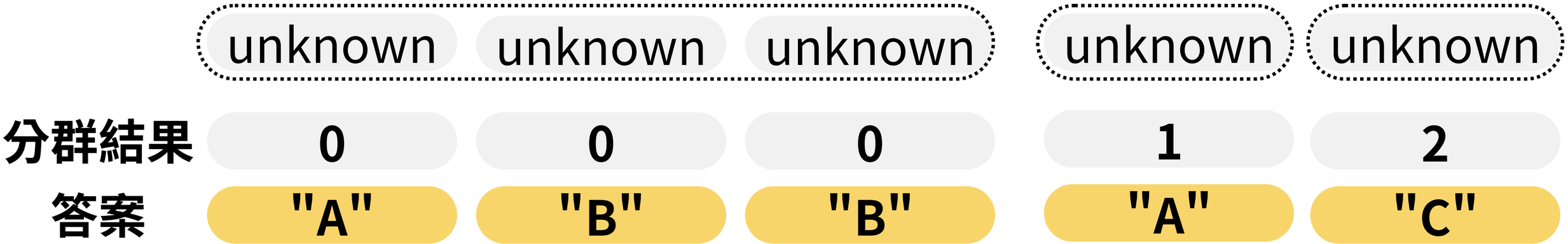
程式流程
STEP 4

無法知道分群所代表的類別，
因此透過投票方式找出對應類別，以此計算準確率。

	unknown	unknown	unknown	unknown	unknown
分群結果	**0**	**0**	**0**	**1**	**2**

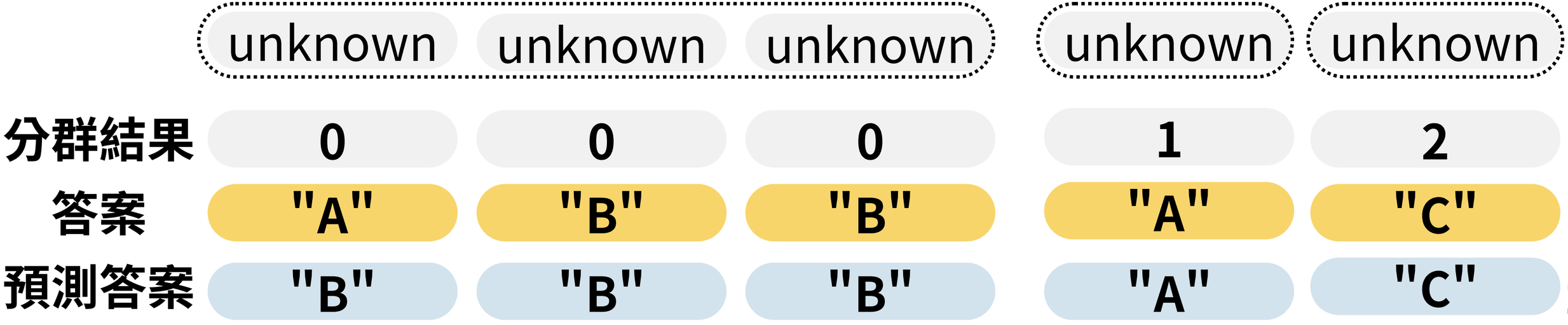
程式流程
STEP 4

無法知道分群所代表的類別，
因此透過投票方式找出對應類別，以此計算準確率。



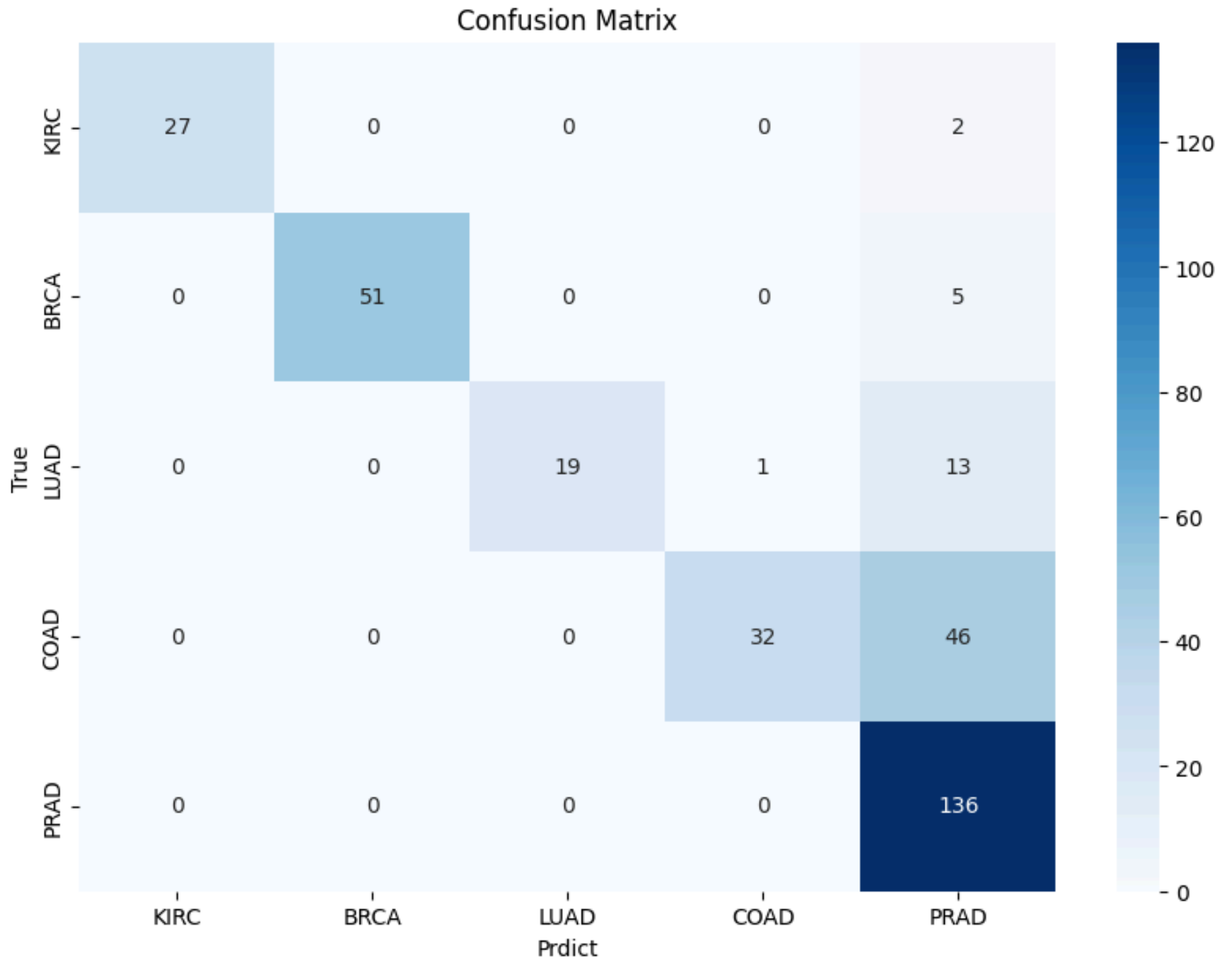
程式流程
STEP 4

無法知道分群所代表的類別，
因此透過投票方式找出對應類別，以此計算準確率。



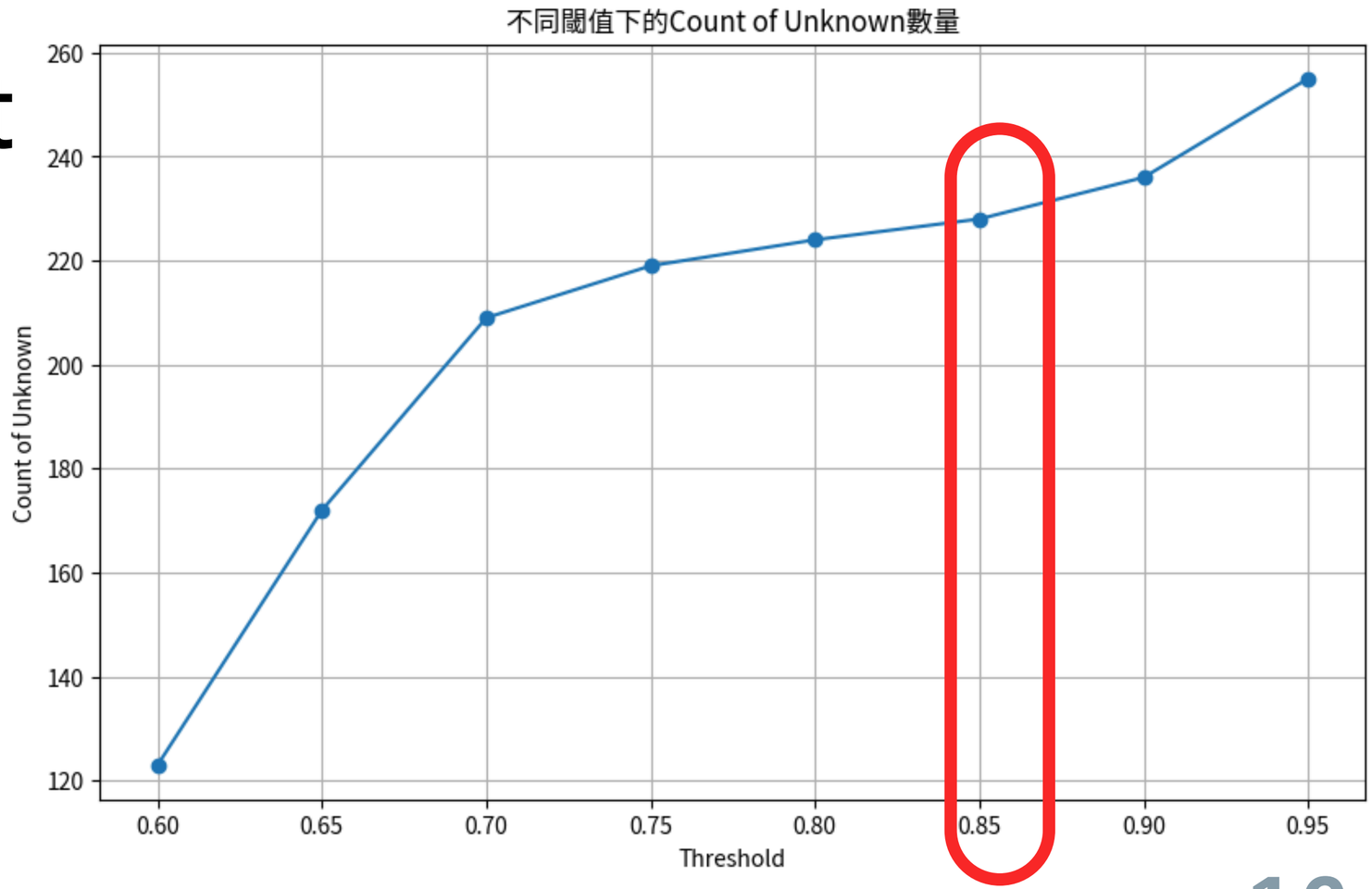
程式流程
STEP 4

Accuracy =
0.7981927710843374



提升準確率

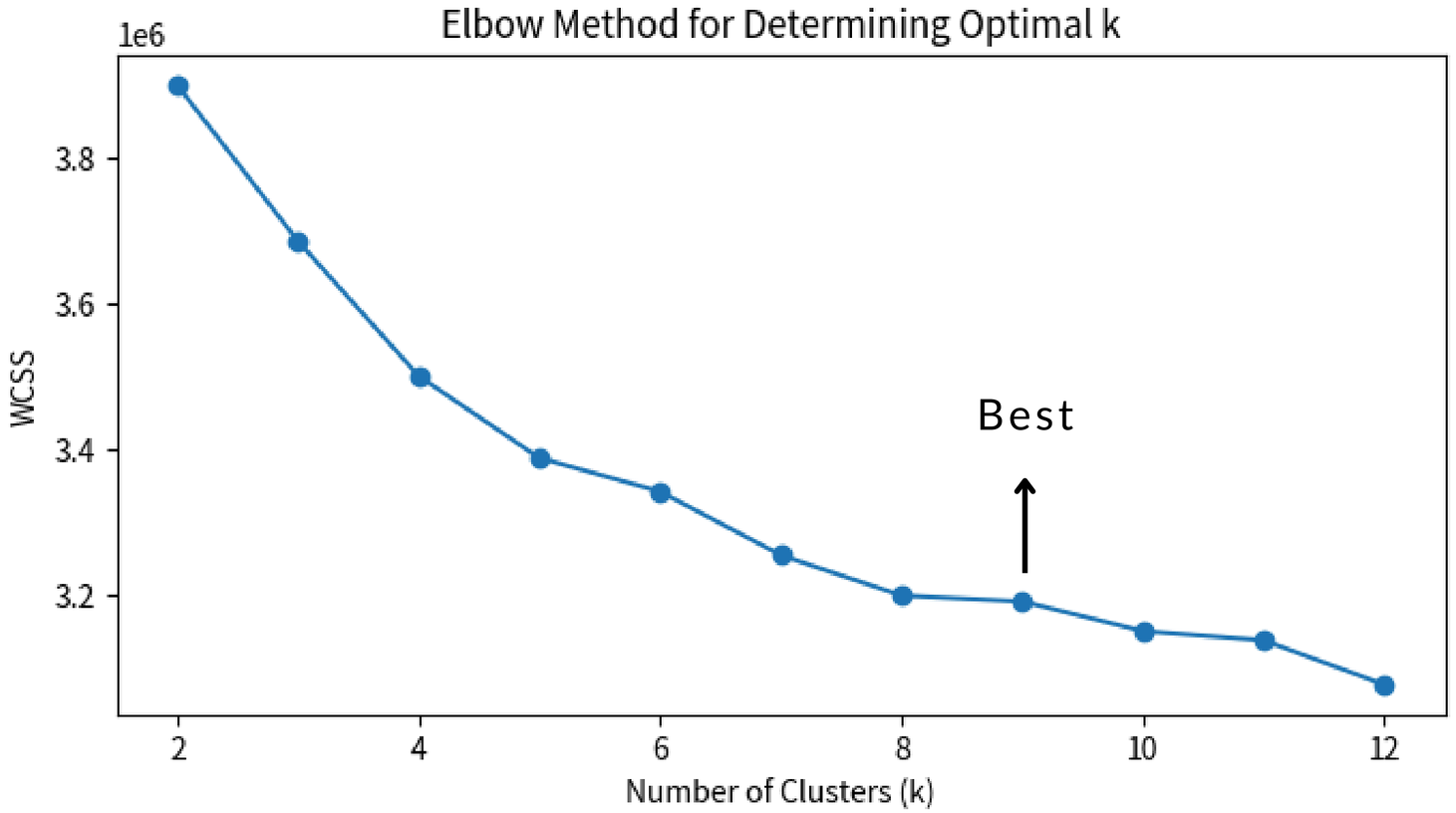
Classification: Random Forest



Clustering Kmeans

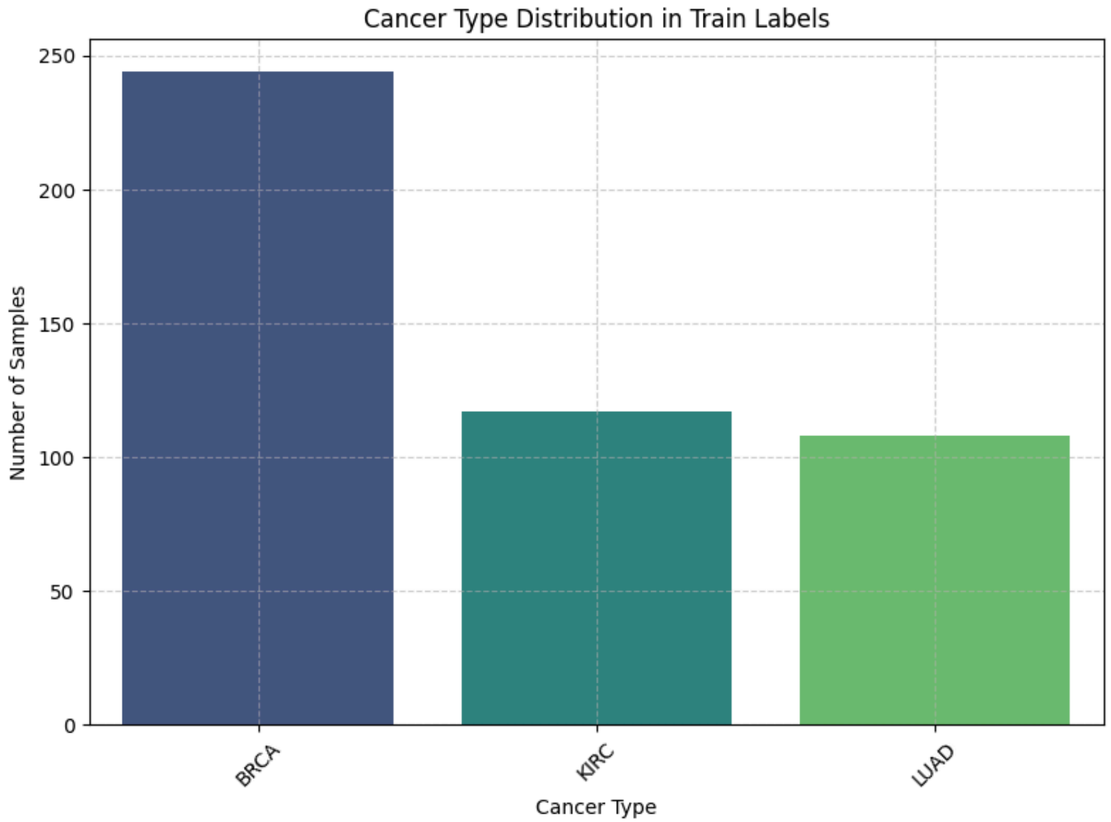
1. 設定要給K個群心

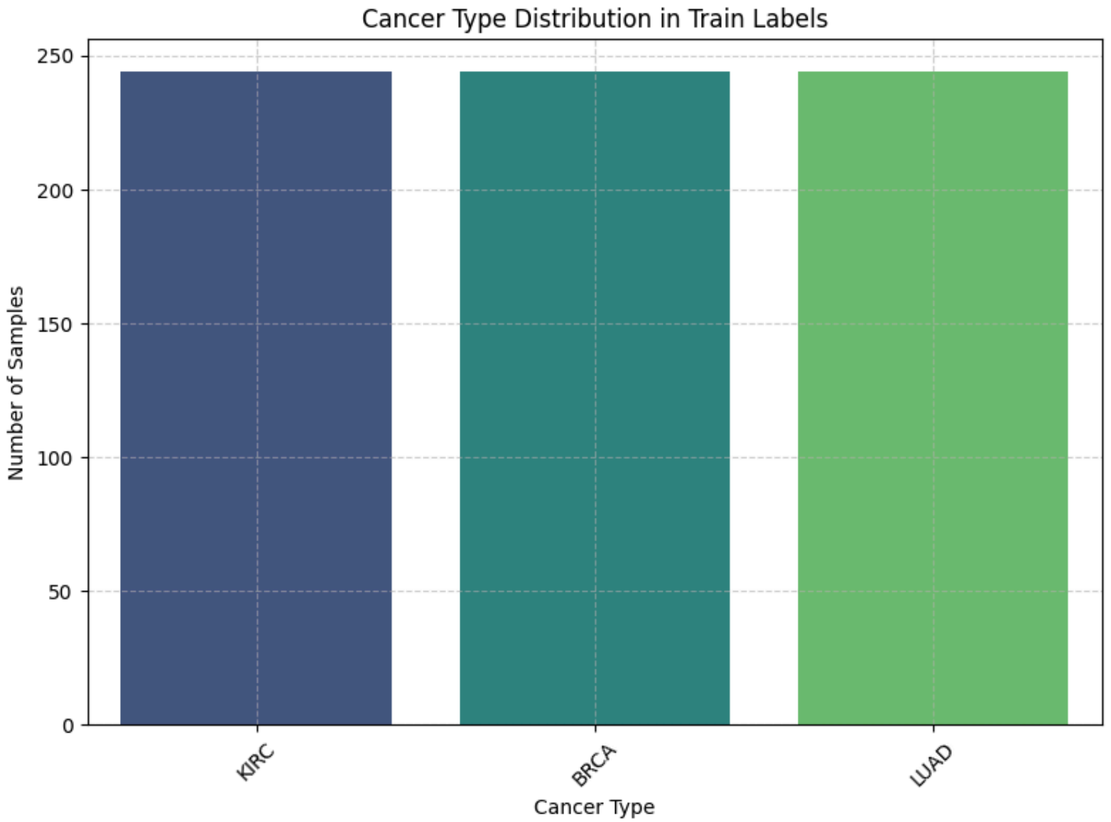
Elbow Method



僅是一種判斷方式，實際測試發現WCSS和最終準確率無關

資料擴增





[469 rows x 20532 columns]

[732 rows x 20532 columns]

資料擴增

01 SMOTE

透過在少數類別樣本之間線性插值來創造新的合成樣本，增加少數類別的樣本數。

02 bSMOTE

是 SMOTE 的一種變體，專注於為邊界上的少數類別樣本生成合成數據。

03 ADASYN

基於 SMOTE 的一個改進方法，其生成合成樣本的數量會根據每個少數類別樣本的學習難度自適應調整。

04 ROS

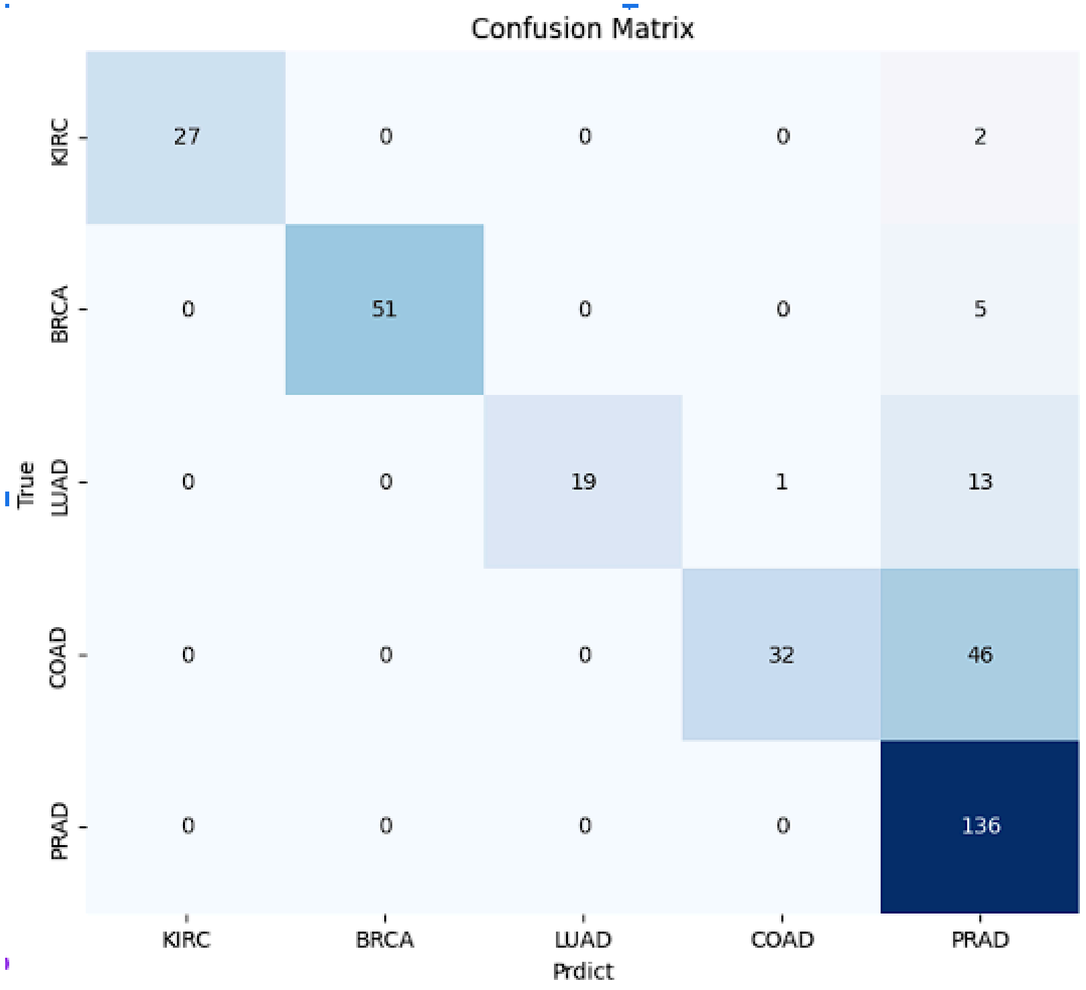
隨機過採樣（ROS）是一種基本的技術，隨機複製少數類別中的樣本來增加其數量。

Result

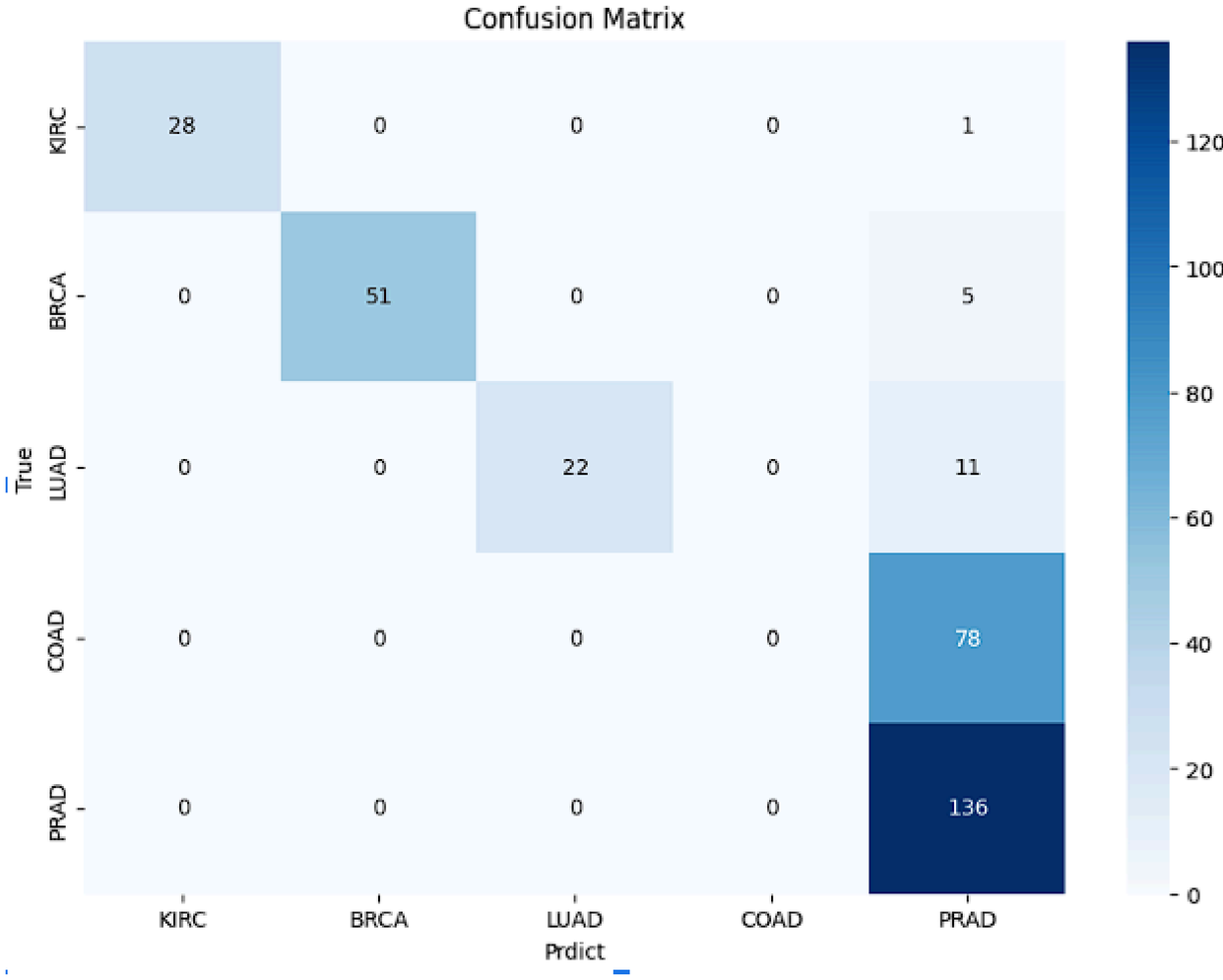
Threshold = 0.85、k_clusters = 9

數據增強方法	準確率 (Accuracy)
一般資料集	**0.7981927710843374**
SMOTE	**0.713855421686747**
bSMOTE	**0.7108433734939759**
RandomOverSampler	**0.608433734939759**
ADASYN	**0.536144578313253**

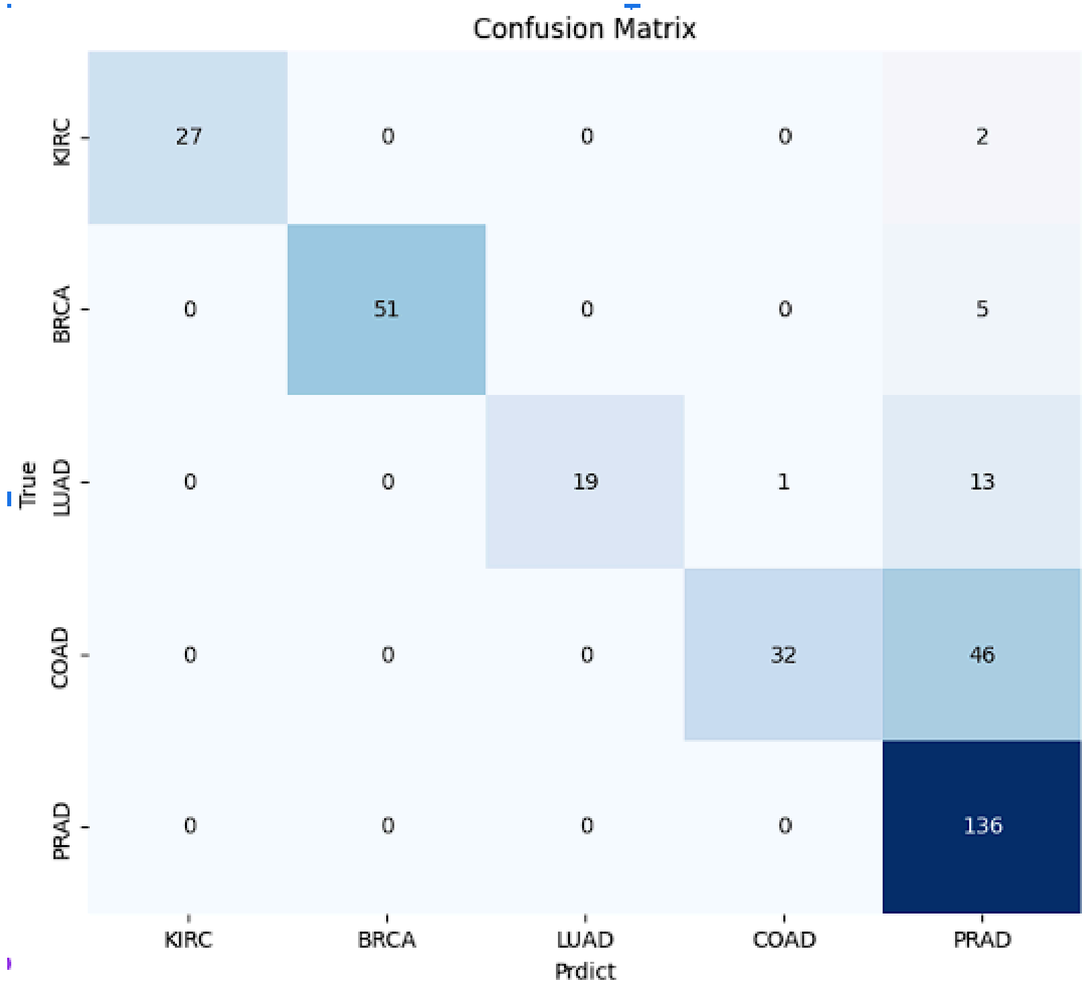
原始資料集



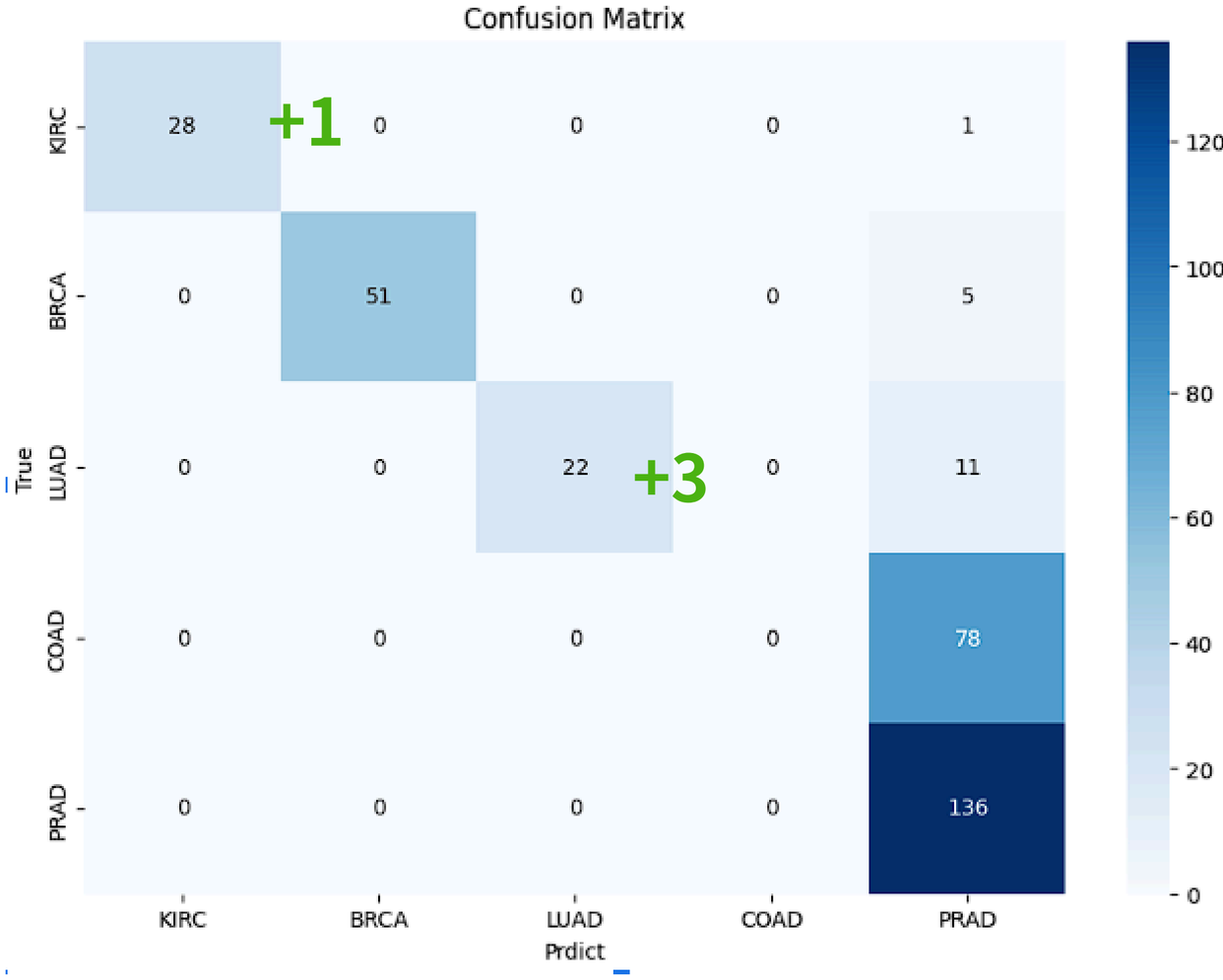
SMOTE 資料集



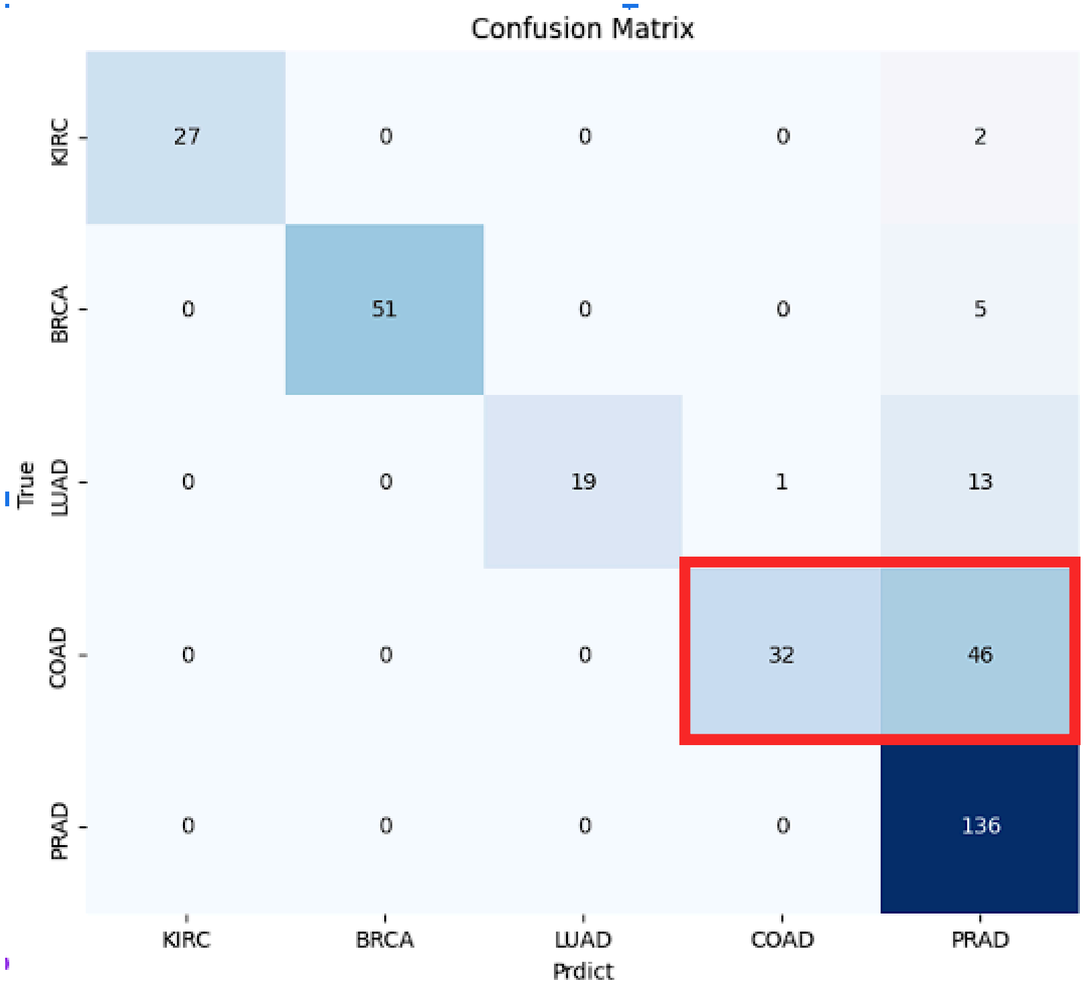
原始資料集



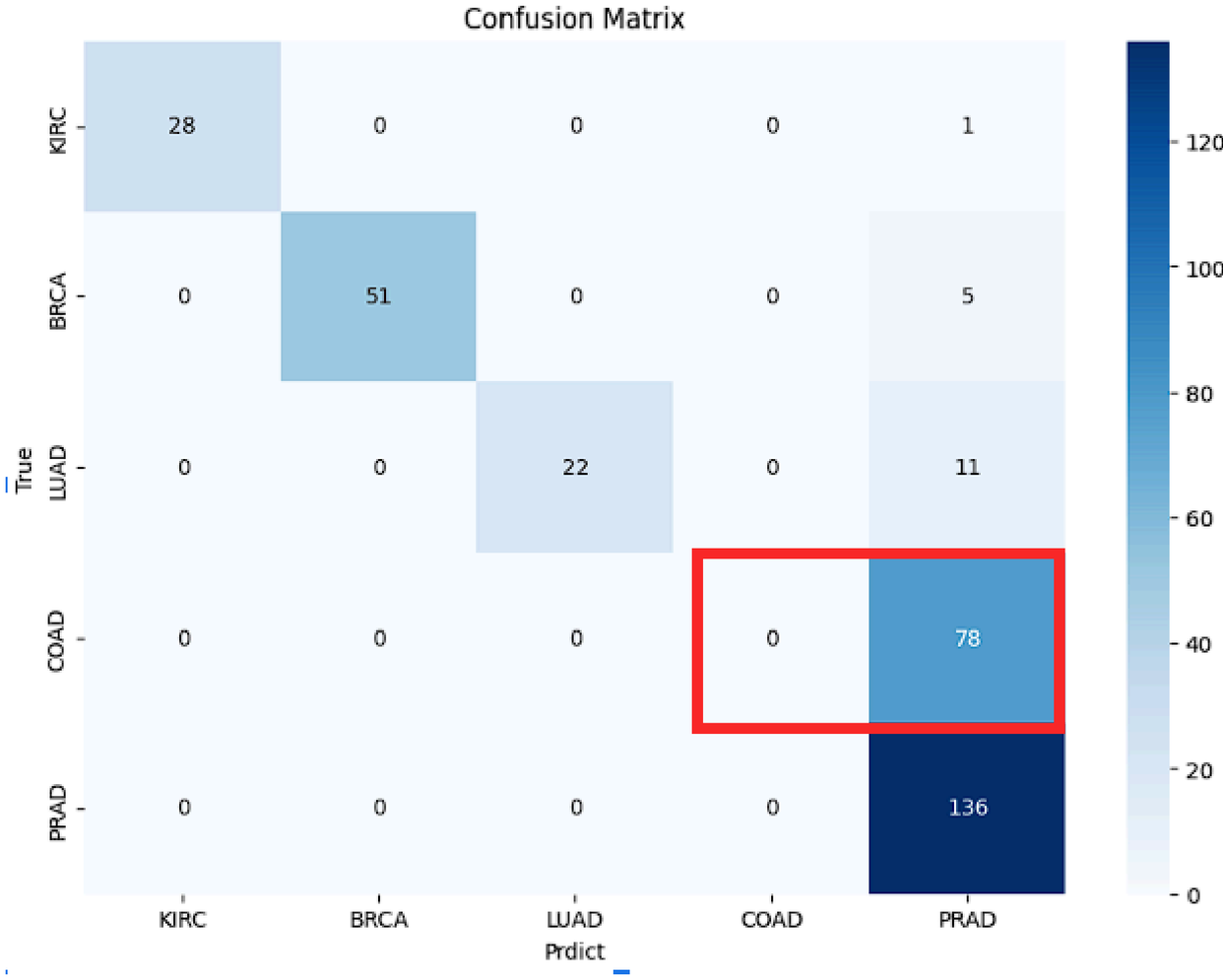
SMOTE 資料集



原始資料集



SMOTE 資料集



Conclusion

Threshold 和 k_clusters 的數值對模型準確率有重大影響

數據擴增能提升"已知"類別的識別能力，但面對未知類別反而容易 overfitting，從而導致準確率下降。