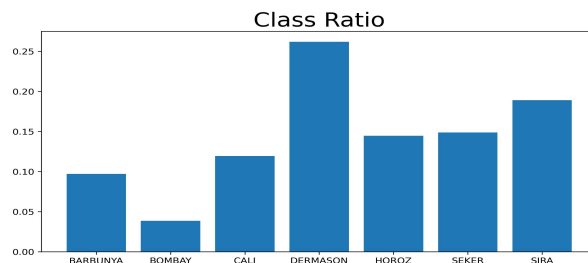


乾豆品種分類器

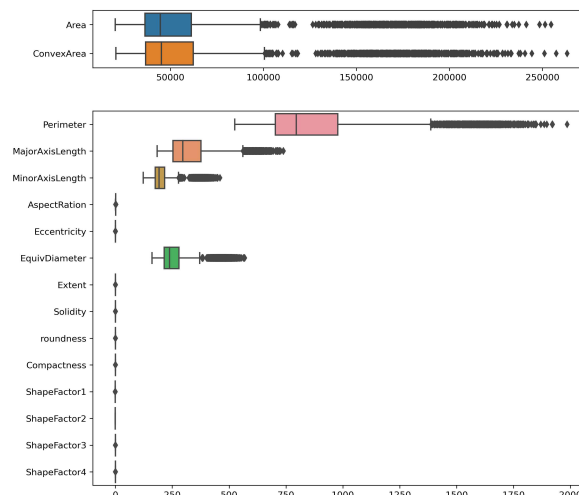
本次研究的資料集——「Dry Bean Dataset」，取自 Kaggle 網站 (<https://reurl.cc/zZoOza>)。干豆在食用豆類中有很高的產量，並具備多個種類，對其種子進行分類可在營銷或生產上發揮作用。該資料集自 7 種不同干豆的圖像中挑選 16 個特徵，其中 12 個為尺寸特徵，4 個為形狀特徵。於是資料集的欄位共 17 個 (含類別)，分別意義如下：(面積意指像素)

1. Area (A) - 區域面積
2. Perimeter (P) - 邊界周長
3. Major axis length (L) - 邊界取兩點的最長線段距離
4. Minor axis length (l) - 垂直 Major axis 的線段距離 (端點在邊界上)
5. Aspect ratio (K) - 長寬比 L/l
6. Eccentricity (Ec) - 偏心率
7. Convex area (C) - 包含種子的最小凸多邊形面積
8. Equivalent diameter (Ed) - 與種子面積相等的圓半徑
9. Extent (Ex) - 邊界框像速與種子面積的比值
10. Solidity (S) - 凸殼中像速與豆中像素的比值
11. Roundness (R) - $4 \cdot \pi \cdot A / (P^2)$
12. Compactness (CO) - Ed/L
- (以下的形狀因子並未說明)
13. ShapeFactor1 (SF1)
14. ShapeFactor2 (SF2)
15. ShapeFactor3 (SF3)
16. ShapeFactor4 (SF4)
17. Class - 類別

該資料集共 13611 筆資料，檢查得知該資料集無缺失值，將資料集以 3:1 拆分為訓練集與驗證集。首先針對訓練集討論，所有欄位皆為數值型 (從特徵說明也能看出)，除 Area、ConvexArea 為整數，其餘特徵為浮點數。以下展示各類別比例，並無嚴重偏態情形。

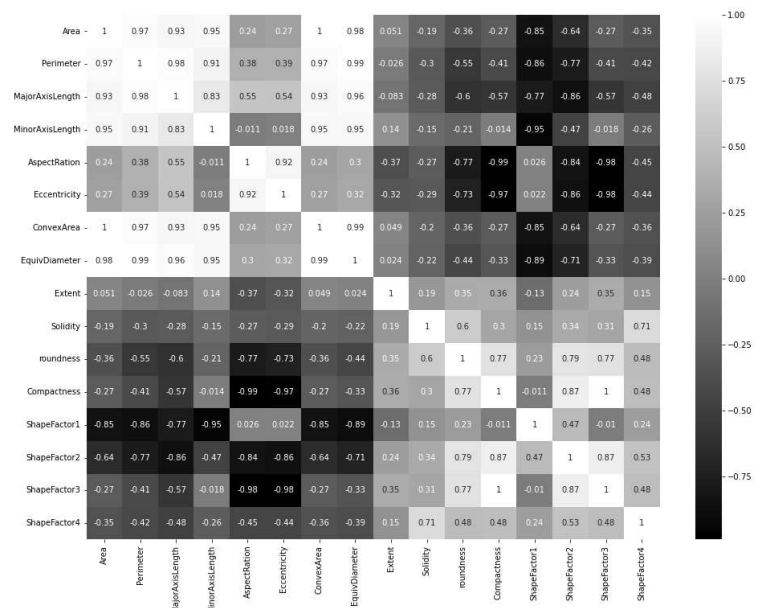


從欄位箱型圖可見各欄位的量級相差甚多。

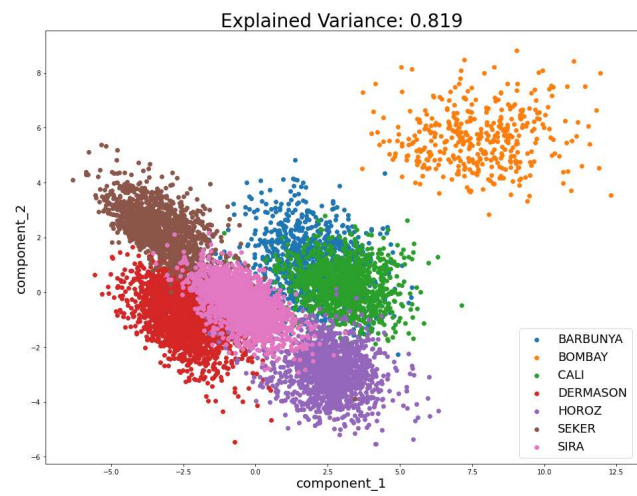


過大的量級差異可能造成權重失衡，於是將資料標準化。由於數據量不大，之後可評估多種分類器。

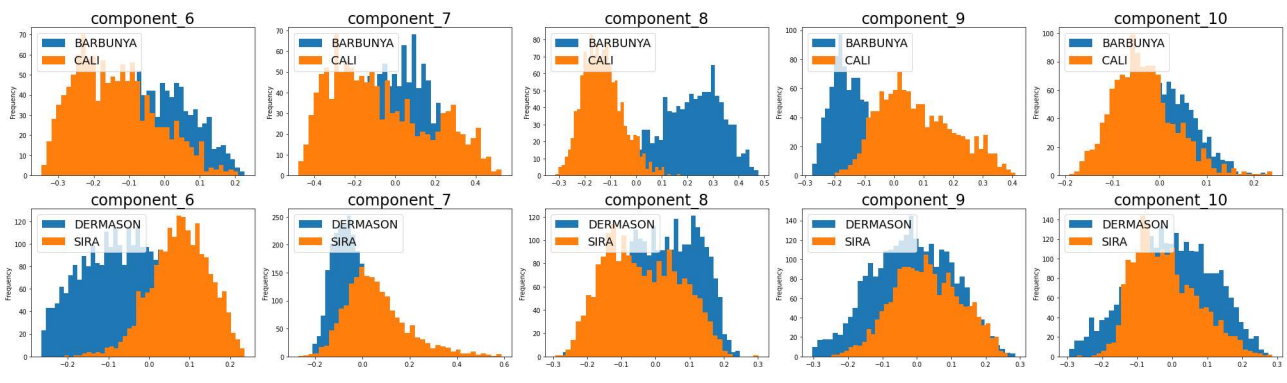
接著事先評估不同類別的分類難度。由於各欄位皆定義自種子的尺寸或形狀，故而在存在相依性。以下繪製欄位間的 Pearson 相關係數。



可見不少欄位相互為高度線性相關 (大於 0.8 或小於 -0.8)，這揭示降維的可能。首先以 PCA 降為 2 維，繪製如下。



可見僅是降至 2 維便保留 81.9% 的變異度，而圖中重疊較多的類別為 (BARBUNYA, CALI)、(DERMASON, SIRA)，即這兩對較難線性區分。另外，將資料先映射至無窮維空間(映射對應 kernel 為 radial basis function)，再以 PCA 降至 10 維，繪製分量 6~10 的分布如下 (分量 1~5 區分不大)。



唯見 (BARBUNYA, CALI) 在分量 8 稍有區分，而 (DERMASON, SIRA) 仍舊不易區分。接著建構分類器。

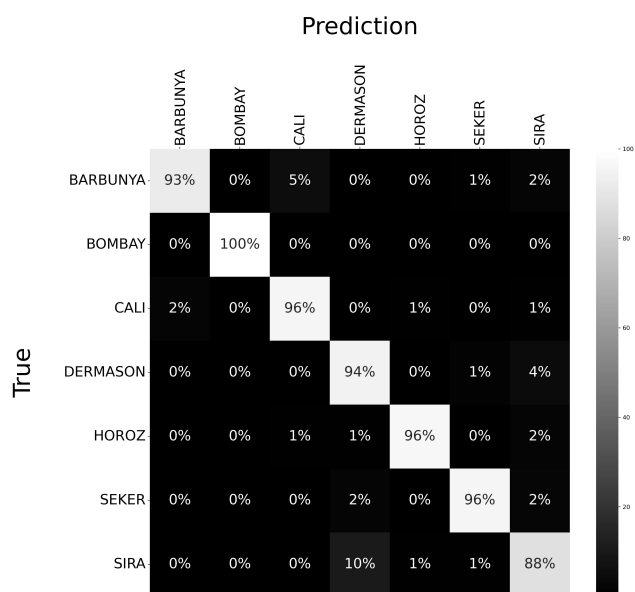
使用 sklearn 的 LogisticRegression、LinearSVC、SVC、GradientBoostingClassifier，分別為羅吉斯回歸、線性 svm、以 radial basis function 作為 kernel 的 svm、梯度提升決策樹。四者皆做交叉驗證，評分採準確度 (accuracy)，結果如下。

分類器	score1	score2	score3	score4	Average
LogisticRegression	0.913	0.929	0.927	0.926	0.924
LinearSVC	0.909	0.920	0.920	0.920	0.917
SVC	0.923	0.930	0.931	0.928	0.928
GradientBoostingClassifier	0.914	0.928	0.926	0.924	0.923

SVC 在每次 CV 都略微勝出，於是採用此分類器，接著以網格搜尋優化參數。以下針對 L2 正則化因子 C 與 kernel 參數 γ 優化，其中 C 愈小則正則化愈高，而 γ 愈小則每個實例的影響範圍愈大。網格搜尋的結果如下：(分數為所有 cv 平均結果)

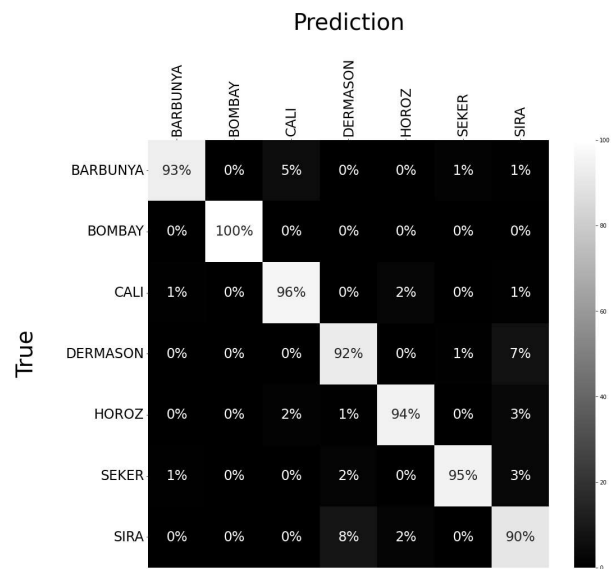
$\gamma \setminus C$	0.1	1	10
0.0200	0.918	0.925	0.928
0.0625	0.922	0.928	0.929
0.1000	0.922	0.928	0.928

最佳模型為 SVC (C=10, $\gamma=0.0625$)，不過各種配置表現相去不遠。接著查看該模型對各類別的判別精度，以下繪出訓練集代入得到的混淆矩陣。



該矩陣呈現每一類別被分類至各類別的比例。顯著的是，SIRA 有 10% 機率被錯分為 DERMASON，反之有 4%；而 BARBUNYA 有 5% 機率被錯分為 CALI。可以預期如此結果，因分類器與第二次 PCA 使用同樣 kernel，而這兩對類別的高度重疊已顯示於先前的 PCA 分析中。

進入驗證階段，將驗證集代入分類器，得到以下混淆矩陣。



同樣地，驗證結果也顯現先前所述的錯分問題，但不嚴重，類別錯分率皆低於 10%。如欲改善，必須為資料集納入其他足以區分這些類別的特徵。