

圖形識別

曾宏鈞 06160485

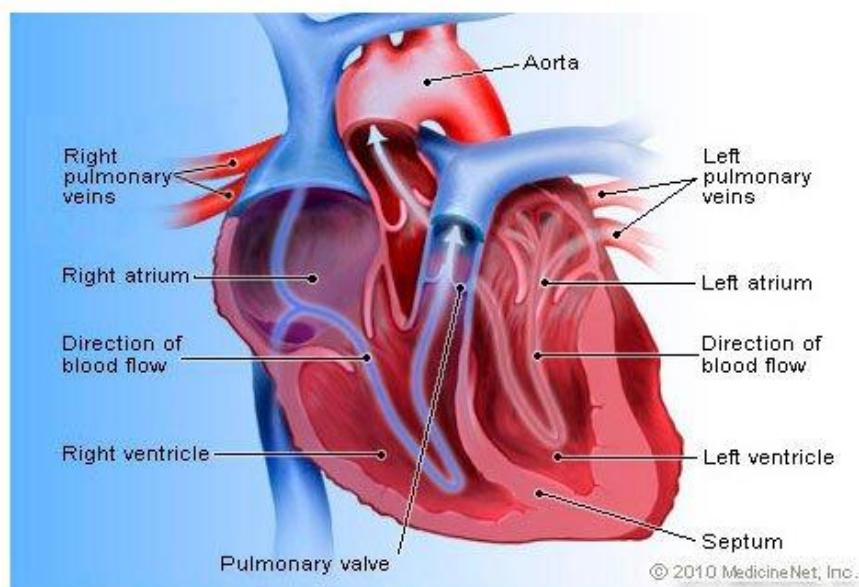
作業二 :tree

資料集：

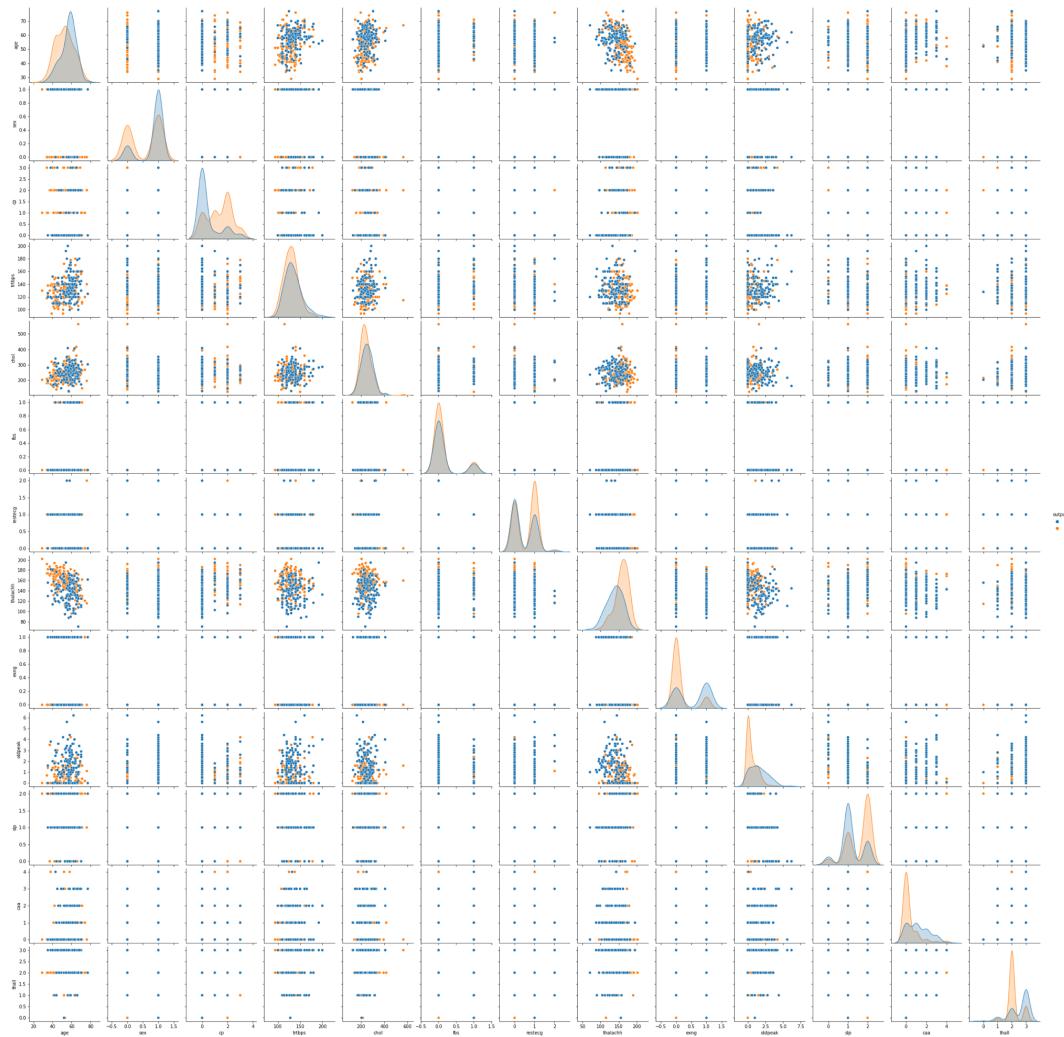
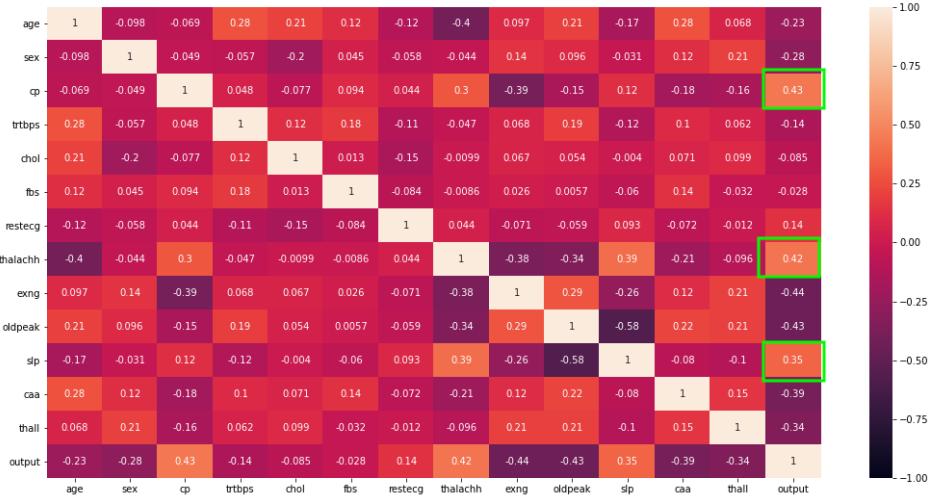
<https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>

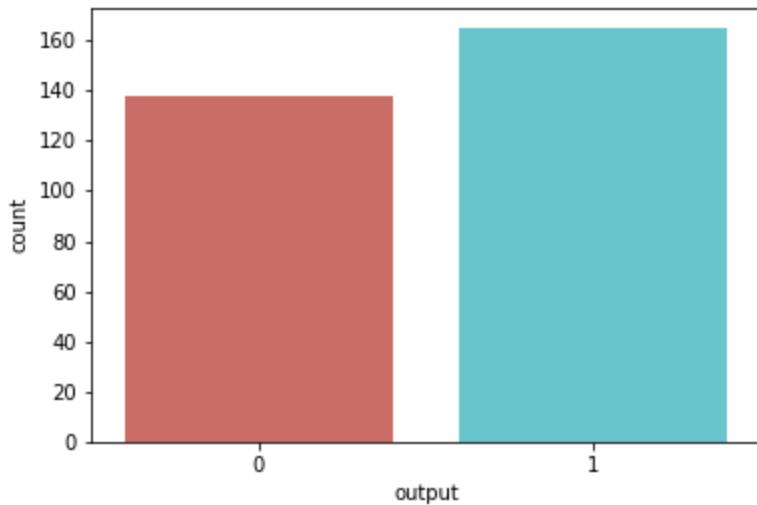
資料

- age: 病患年紀
- sex: 病患性別
- cp: 胸痛的類型, 0 = 非典型心絞痛, 1 = Atypical Angina, 2 = 非心絞痛, 3 = 無症狀
- trtbps: 靜止時血壓 (in mm Hg)
- chol: BMI sensor 測量的膽固醇(mg/dl)
- fbs: (空腹時的血糖 > 120 mg/dl), 1 = True, 0 = False
- restecg: 靜止時的心電圖結果, 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
- thalachh: 最大心率
- oldpeak: 上一個高峰
- slp: 斜率
- caa: 血管數
- thall: Thalium Stress Test result ~ (0,3)
- exng: 運動誘發的心絞痛 ~ 1 = Yes, 0 = No
- output: 目標



資料探勘

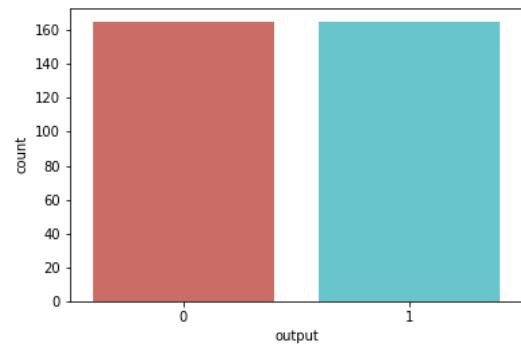
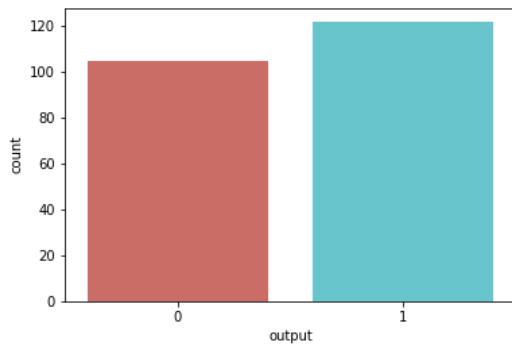




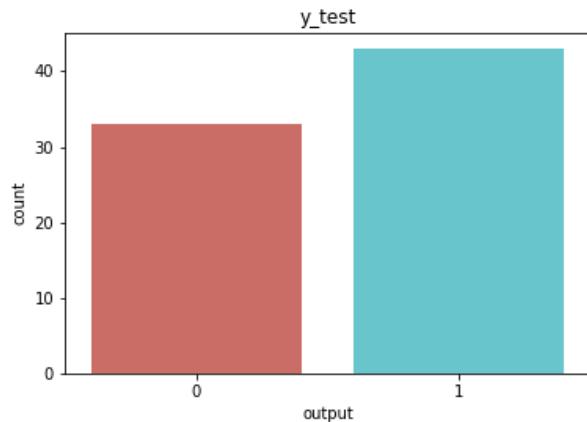
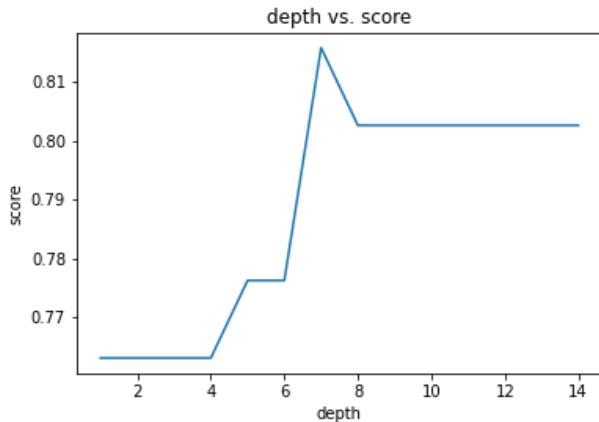
- 觀察上圖，資料相關性與特徵 cp, thalachh 和 slp 相關性較高。
- 類別之間沒有明確的線性相關
- 在pairplot中可以看出有一些離群值
- 原始資料的目標(output)沒有平衡

資料工程

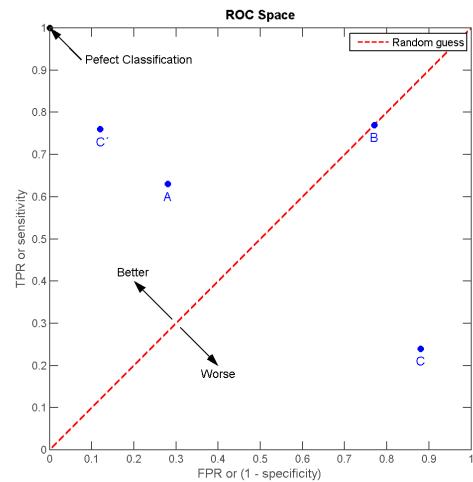
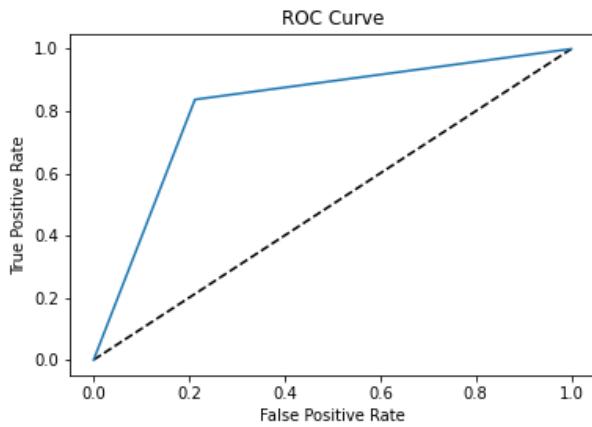
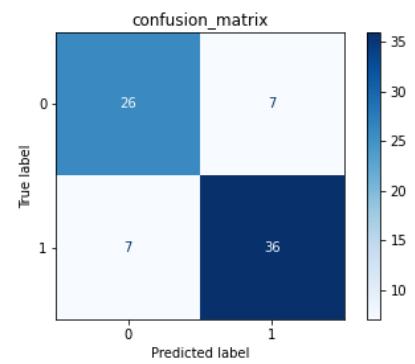
SMOTE 平衡資料



模型分析



	precision	recall	f1-score	support
0	0.79	0.79	0.79	33
1	0.84	0.84	0.84	43
accuracy				
macro avg	0.81	0.81	0.81	76
weighted avg	0.82	0.82	0.82	76

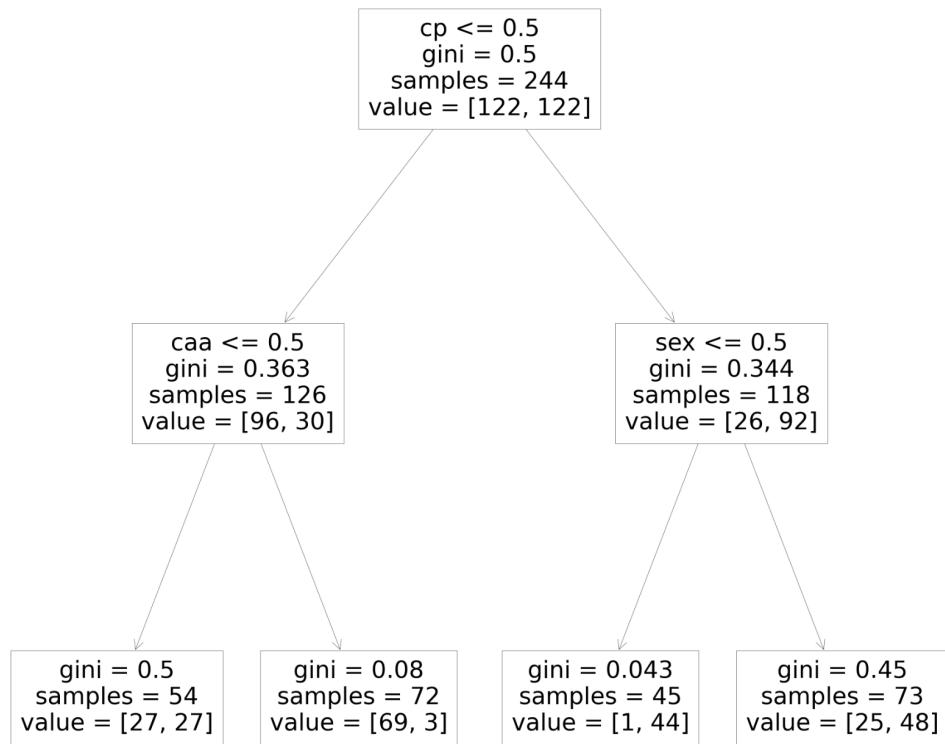


- 模型在深度(depth) = 7時準確度最高，太深準確度停留在0.80是因為葉節點無法再繼續計算gini(亂度為0)，觀察img/*tree_viz.png
- 原始資料沒有平=>使用SMOTE方法平衡資料
- 1的類別比0的類別準確度高(原本的y_test 1本來就比較多)
- Confusion matrix
 - Accuracy: 81.5
 - Recall(TPR): 0.83
 - Precision: 0.837
 - F1-score: 0.837
 - 完整的cm在檔案夾...
- 因此，若使用此模型進行心臟病的預測，總共76個病例裡，有36人確實為有心臟病(TP)、26人為沒有(TN)、檢測錯誤(FN)及(FP)各7人，可以透過多檢測幾次使病人確定是否為心臟病。
- 從ROC曲線可以看出，此模型優於隨機猜測，妥善調整閥值可以獲得好的預測。(AUC:0.81)

	feature	importance
2	cp	0.293
11	caa	0.110
0	age	0.097
7	thalachh	0.094
12	thall	0.085
9	oldpeak	0.069
4	chol	0.065
1	sex	0.060
3	trtbps	0.054
5	fbs	0.033
6	restecg	0.015
10	slp	0.015
8	exng	0.008

- 以特徵重要性而言，重要的是cp(疼痛的類型)、caa(血管數目)，可以從這兩點去探討心臟病與這兩個特徵之間的關係。

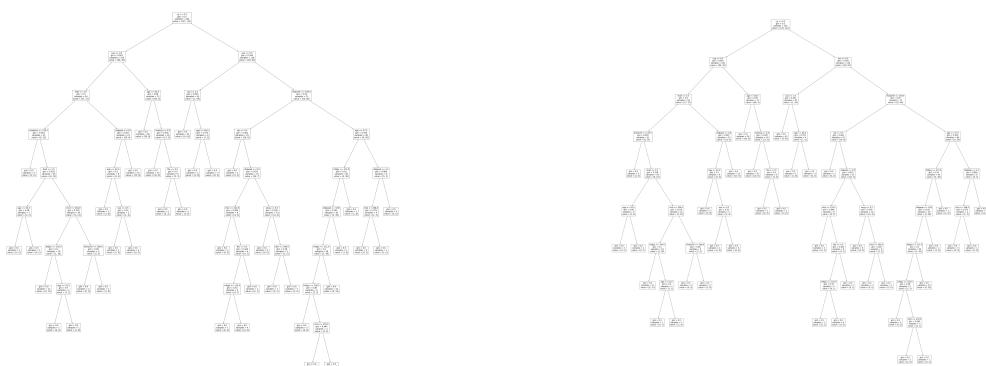
視覺化訓練的決策樹



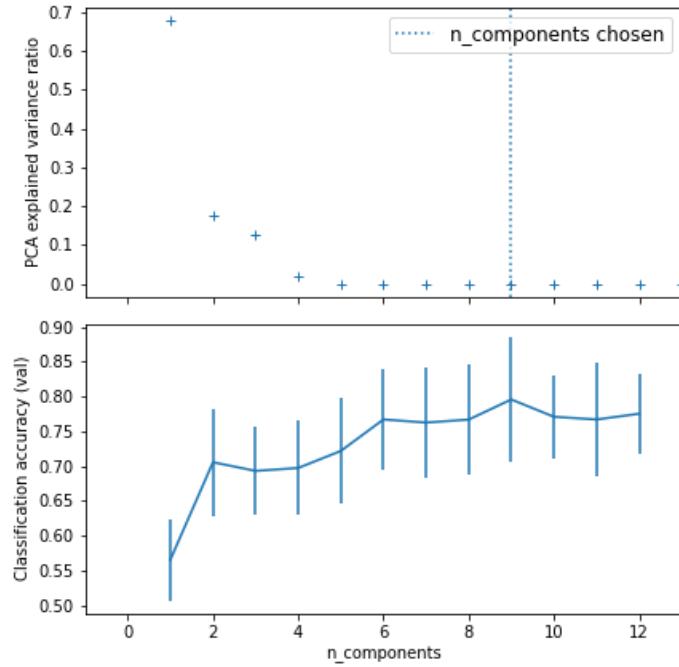
最大深度為2時，樹的圖

完整的圖檔在檔案夾...

- Root為 $cp \leq 0.5$, true: 左邊, false: 右邊...
- 在深度9以後圖型一樣(葉節點亂度0, 無法再繼續細分), 下圖為深度9、12



使用GridSearchCV及PCA找最佳值



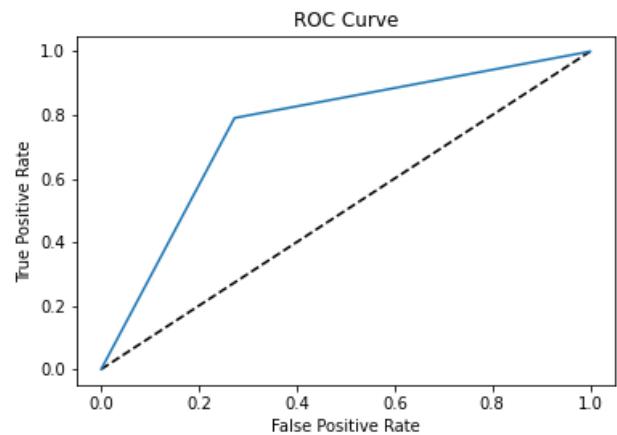
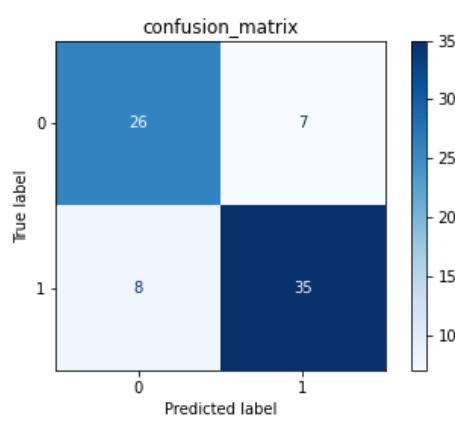
- 由上圖可知，在PCA取的特徵向量9(14(原本)->9)時，準確度最佳，此時最佳的樹深度為7
- 做完PCA後，雖然維度縮減到9，但比起原本準確度沒有上升。

使用

precision recall f1-score support

0	0.73	0.73	0.73	33
1	0.79	0.79	0.79	43

accuracy		0.76	76	
macro avg	0.76	0.76	0.76	76
weighted avg	0.76	0.76	0.76	76



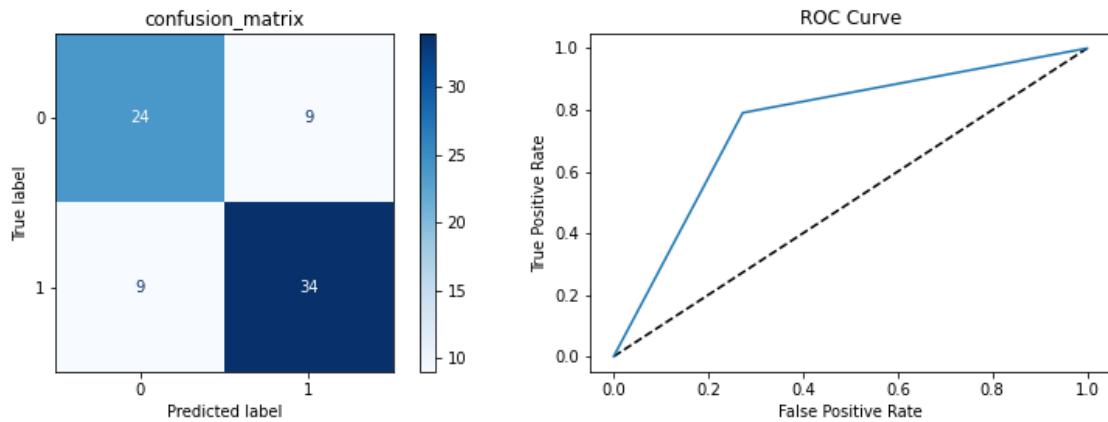
使用GridSearchCV找尋其他參數

- Max_depth:樹深 [1,15]
- Criterion : 使用的標準 ['gini', 'entropy']
- max_leaf_nodes: 葉節點最大限制 [2,4,6,8,16,32,64,128]
- Min_samples_split: 內部節點元素最小需拆分的數目 [2, 3, 4]

我們使用以上參數，使用cross validation(cv=5)的方式使用GridSearchCV搜尋最佳參數，得到最佳的cross validation score為0.8，參數如下：

Best parameter (CV score=0.804):

```
{'tree_criterion': 'entropy', 'tree_max_depth': 3, 'tree_max_leaf_nodes': 8,
'tree_min_samples_split': 3}
```



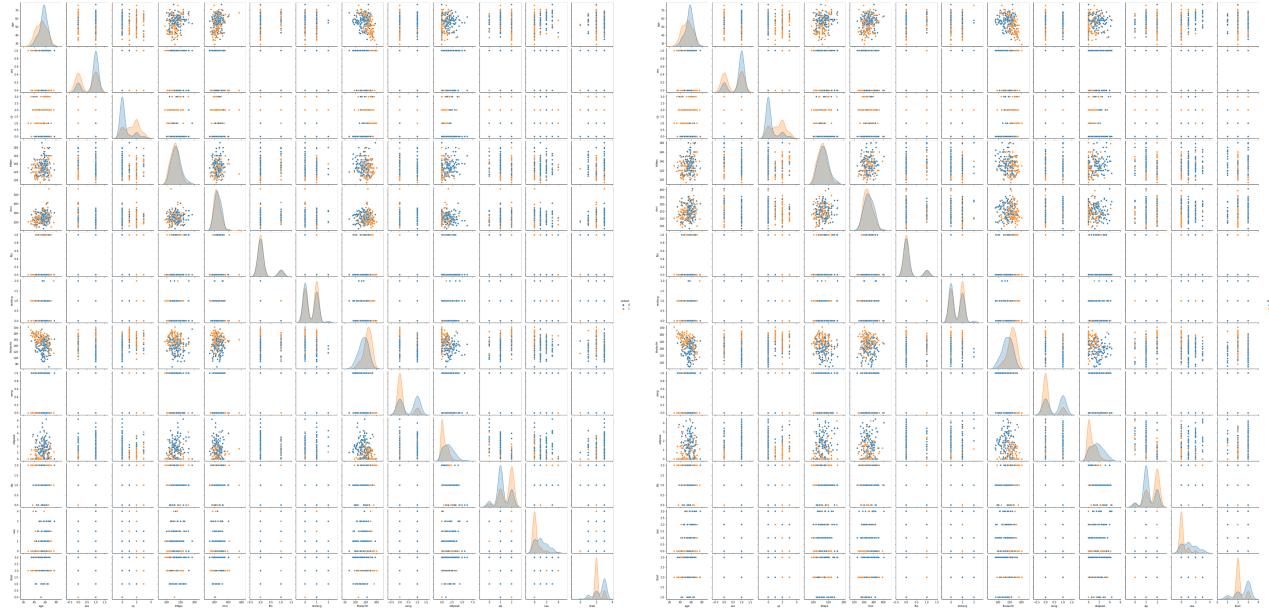
	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.73	0.73	0.73	33
1	0.79	0.79	0.79	43

	accuracy			
--	----------	--	--	--

macro avg	0.76	0.76	0.76	76
weighted avg	0.76	0.76	0.76	76

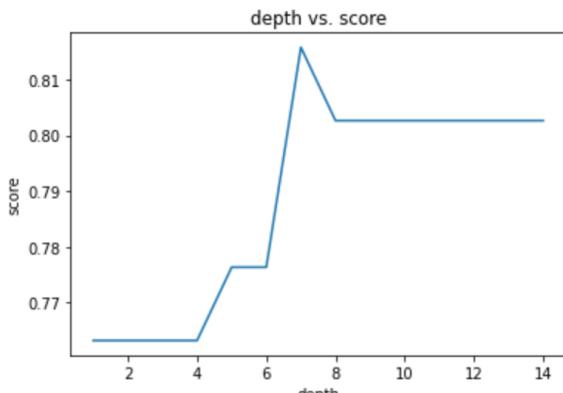
去除離群點



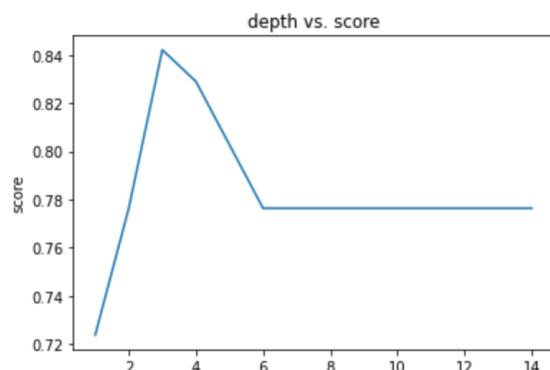
原始訓練(左), 移除outlier(右)

- 訓練集pairplot分析可知有離群點
- 訓練時去除Outlier可以使準確度提升(0.81->0.84)
- 訓練時去除Outlier使決策樹的深度減少(7vs3), 由於不需要細分離群值

best: 0.8157894736842105 depth: 7



best: 0.8421052631578947 depth: 3



使用不同模型分析準確度



發現的問題

Q1:每次使用GridSearchCV出來搜尋的模型參數皆不同