Report

# dataset

採用一個fruits的dataset，裡面包含四種水果，apple、mandarin、orange、lemon。

合計59個item及7個feature

每一種水果有7個feature，

fruit\_label:標籤，每個水果分別有1~4的標籤，apple:1，mandarin:2，orange:3，lemon:4

fruit\_name:名字

fruit\_subtype:各種水果都有品種

mass:重量

width:寬度

height:長度

color\_score:把顏色數值化，判斷是否成熟

分類器

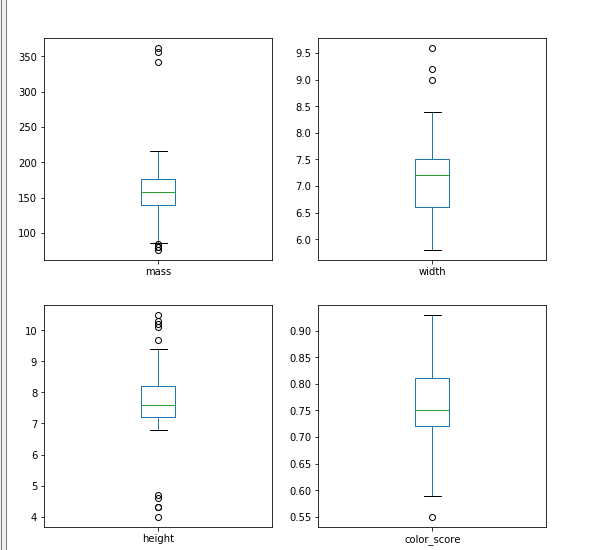
我選擇使用python內建scikit-learn，利用裡面的分類器，去找出最好的分類器

Implement

fruits.drop('fruit\_label', axis=1).plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False, figsize=(9,9),   
 title='Box Plot for each input variable')

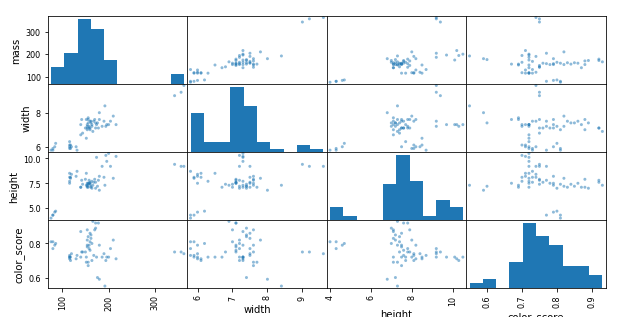
plt.savefig('fruits\_box')  
plt.show()

### 把一些feature Visualization



看起來顏色的分數比較接近Gaussian distribution

接著把feature整合一下



接著就可以使用機器學習開始training了

以下是幾個分類器所比較出來的結果

Logistic regression



Decision tree

第一個不設限，第二個則是限制高度為3



KNN



#### Linear Discriminant Analysis



Gaussian Naïve Bayes



SVM



結論:

如果依據理論來看，KNN的訓練效果應該是最好，但此筆DATA最好的結果是Decision tree所訓練出來的，而regression不意外的則是墊底

。