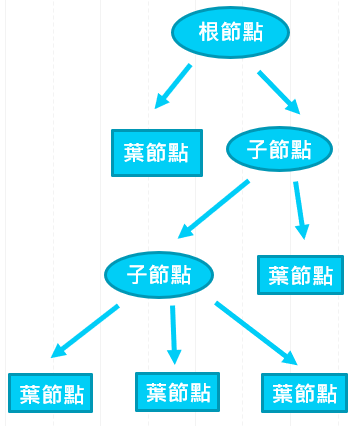
Random Forest Note

* 決策樹(decision tree)

模仿人類做決策的過程。

決策樹的結構由根節點、子節點、葉節點組成一個樹狀的結構。



一開始會先從根節點開始，然後依據各個特徵將資料分割到兩邊。

為了使用好的特徵來做分割(也就是能將數據分得較乾淨的特徵)，我們就需要透過信息增益(information gain)來判斷，如Gini不純度、Entropy。信息增益是特徵選擇的一個重要指標，它定義為一個特徵能能夠為分類系統帶來多少訊息，帶來的訊息越多，說明該特徵越重要。以下為它的公式:

F表示為分割的特徵，代表根節點的資料，為第j個子節點的資料，為該節點的資料數量。，為分類完整的程度。

1. Gini不純度(Gini impurity) : 代表在節點t，屬於類別c的比例。
2. Entropy : 對不確定性的測量
3. 分類錯誤率(classification error)

Gini不純度是基於節點來計算的，樹中的每個節點都會有一個不純度，並且子節點的不純度一定是低於根節點的，也就是說，在同一棵決策樹上，葉子節點的不純度一定是最低的。

Entropy的計算比Gini不純度要緩慢， 這是因為不純度的計算不涉及對數

因為Entropy對不純度非常的敏感，所以以 Entropy作為指標時，決策樹的生長會更加精細，對於高維資料或者噪音很多的資料，Entropy很容易過擬合，這個時候使用不純度效果往往比較好，這也是為什麼不純度比較常見的原因。

其中樹的種類又分很多種， CART(Classification And Regression Tree)是一個很經典的演算法，它的信息增益為搭配Gini不純度使用。

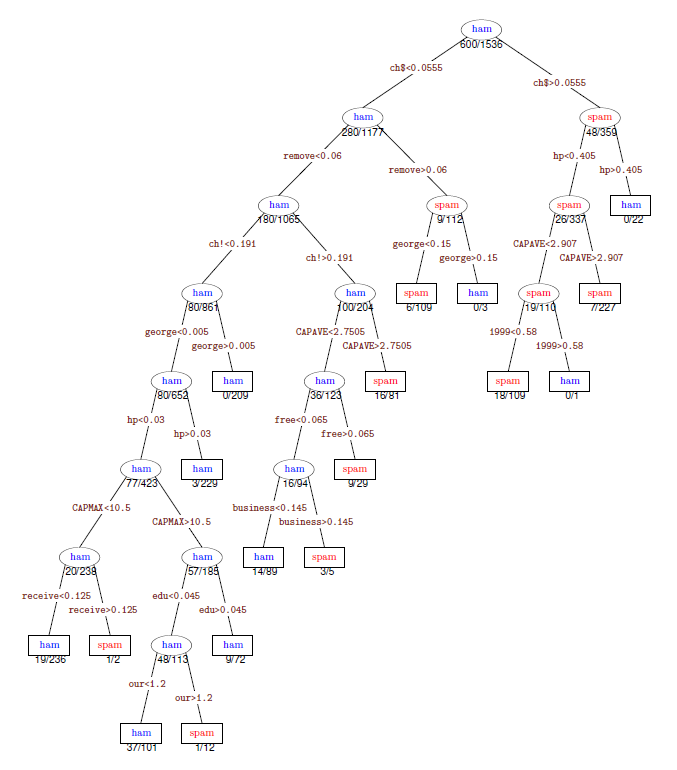
樹的優點 :

1. 樹會自動選擇在模型中用於決定分類的變量
2. 分而治之的演算法
3. 模型直觀，便於理解，應用廣泛
4. 可以處理混合特徵和具遺失的資料
5. 小型的樹很容易去解釋

樹的缺點 :

1. 大型的樹很難去解釋
2. 樹通常預測的不太好
3. 缺少足夠的理論支持

樹的結構越複雜(濃密)，他們的變異數越大。較前面的節點決定架構，後面的節點則透過重要的變數將訓練數據分類到靠近目標的區域，因此後面節點分類的誤差會較小。



* 隨機森林 (random forest)

Random Forest = Bagging + Decision Tree

* 隨機森林生長許多樹，希望透過平均去降低變異數。因此在生長樹的過程加入了一些隨機化，隨機化這裡用兩個方式包括bootstrap sampling 或對觀測值子抽樣(subsampling)，以及對變數的子抽樣，一種集成學習(ensemble learning)技術。

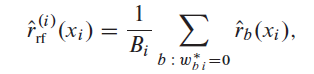
1. Bootstrap : 透過每一次對訓練資料bootstrap sampling的resampled training dataset去生長每一棵樹， 使得每次的樹有一些不同，相關性沒那麼大。
2. 分類變數隨機化 : 每次分類，只選擇p個變數中的m個隨機變數子合。通常選擇.

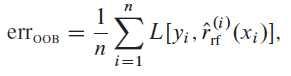
**隨機森林演算法**

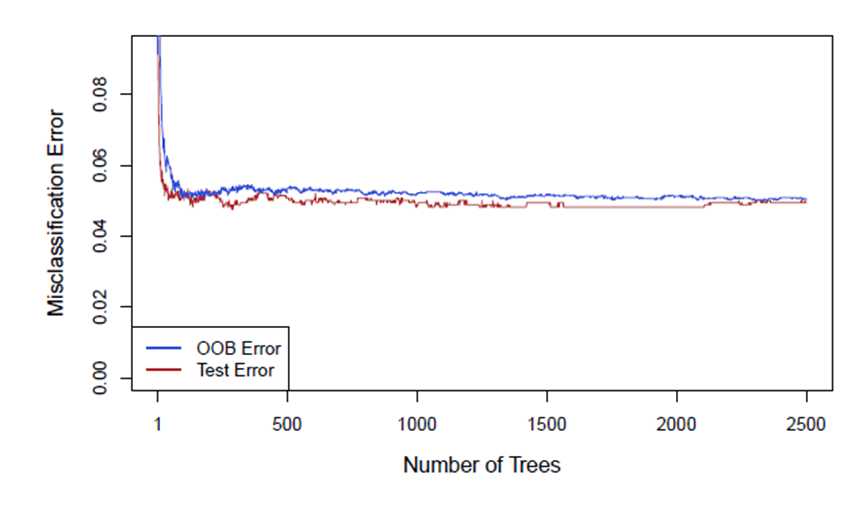
1. 給定訓練資料 =(, ) 。決定分類變數個數m ≦ p和樹的個數B
2. 對每一棵樹 b = 1,…,B :
3. 先在n列訓練資料中隨機重複抽出n列當次訓練資料 (bootstrap)
4. 在每個分類之前隨機採樣m個變數，使用去生成樹
5. 將預測資料帶入隨機森林
6. 計算訓練資料中對沒有被bootstrap抽樣到的資料i的反應變數的OO誤差，整體OOB誤差會是OO的平均。

* Out-of-Bag(OOB) Error Estimates

因為使用bootstrap 生成訓練資料，因此原始訓練資料中有一部份不會出現在訓練資料中，這些資料便稱為OOB資料，對每一棵樹都透過OOB資料去估計誤差，將森林中每一棵樹的OOB誤差取平均，便得到隨機森林的OOB誤差估計。

這個式子是計算森林沒被抽到

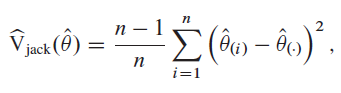
 這個就是用沒被抽到的yi和沒被抽到森林去計算OOB誤差



因為通常用來估計分類器的方式為交叉驗證(Cross Validation)，從這張圖可以看到，相較起交叉驗證，OOB 估計能透過少量資料的計算量達到近似於交叉驗證的結果，對於交叉驗證的高計算量下，是一個節省資源的抽樣及估計方式。

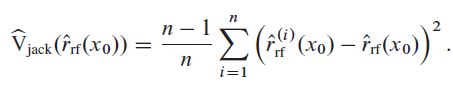
* Standard Error

Jackknife(刀切法)透過切割data來估計偏差及變異數。



為使用除了觀測值i以外觀測值的估計，為的平均。

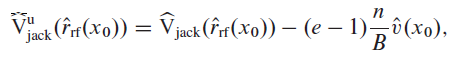
那麼，應用在隨機森林的jackknife表示為 :

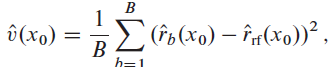


這個式子是在B為無限大的情況下推導的，然而在現實中B是有限的，這個式子會有偏誤和變異數。

因此會希望在估計變異數時，用比在隨機森林時更大的B。

或者可以使用和隨機森林時相同的B，以及jackknife估計的校正版本 :





隨機森林中樹的變異數估計很容易得到

決策邊界附近的估計往往具有更高的標準誤差

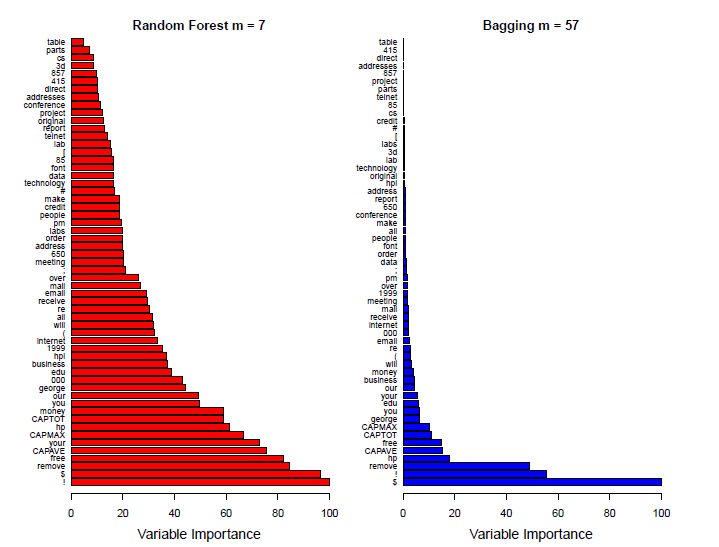
* Variable-Importance Plots

每棵樹用來分類的變數子集雖然不一定會相同但可能會重疊。因此可能會得到一個結論是 : 沒在任何樹中作為分類變數的變數可能都不太重要(不能很好的做分類)。

因此variable-importance plots為一種評估變量相對重要性的方法。

在樹中使用各個變數時，演算法紀錄了因為這次分類導致的分類標準的減少，

在所有樹加總後對每個變數做整理，作為變數相對重要性的度量圖。



這張圖可以看到是在分類變數隨機化的情況下，變數重要性分散的比bagging好，也就是說bagging通常在分類時只選擇分類最好的變數。可以看出在m較小時，變數選擇傾向在相關變數中平均分配。

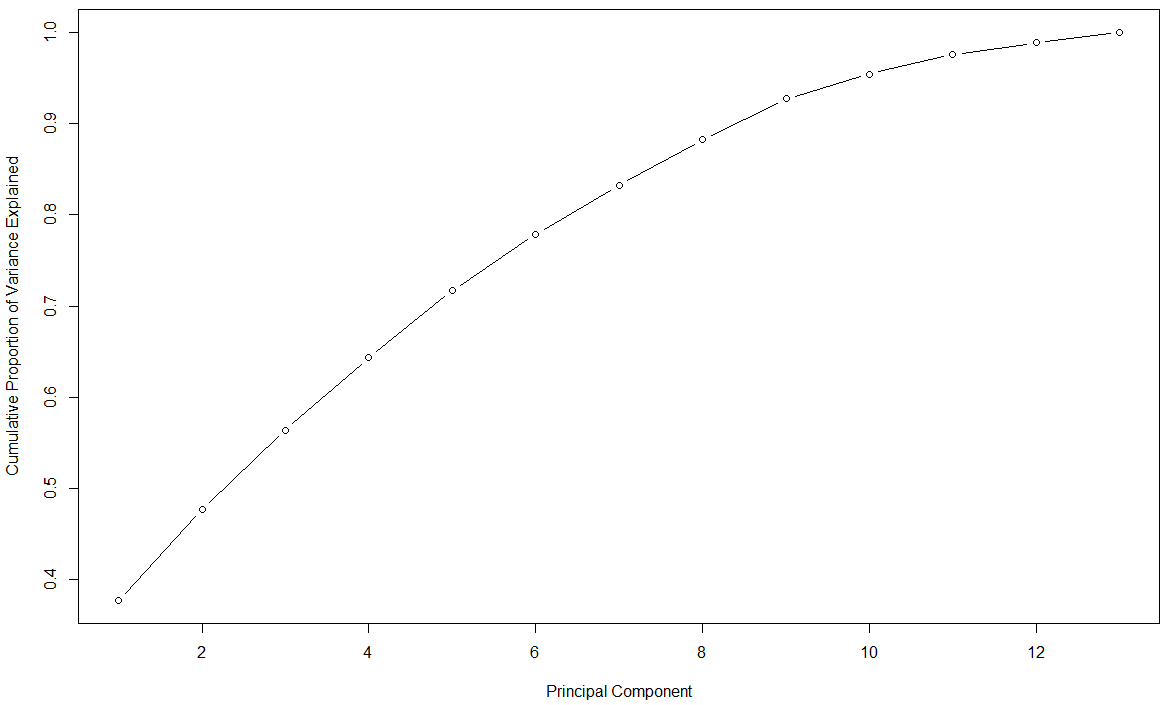
* Breast Cancer Analysis

資料為血液中的微量元素，label分成1、2、3，1為惡性，2為良性，3為健康，目標是探討血液元素與罹患乳癌的相關性。

將資料標準化，拆成80%訓練組、20%測試組

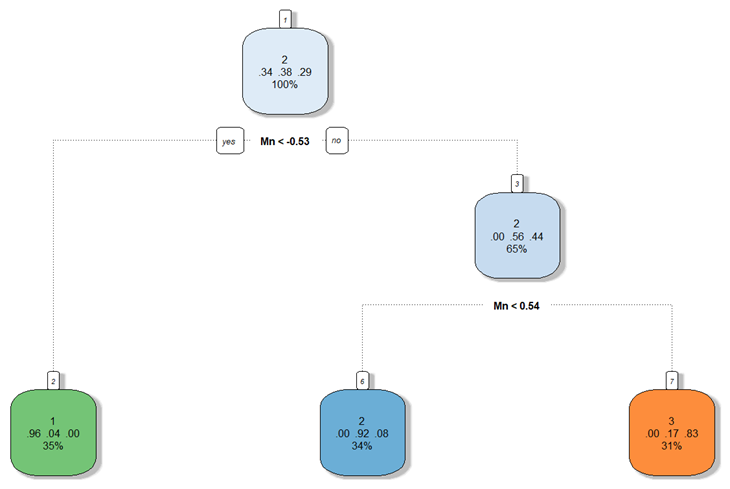
因為bagging關係，所以不再將訓練組細分為訓練組及驗證組

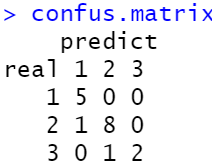
先做PCA查看是否需剔除一些特徵



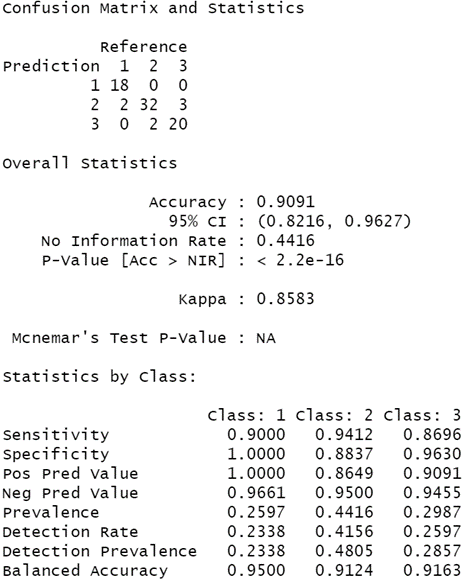
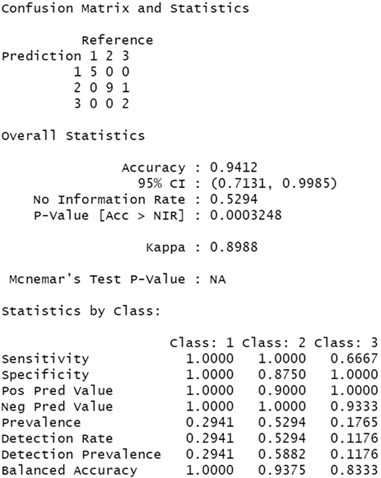
由圖可發現變異解釋在最後並沒有在1的部分有平穩現象，因此使用所有特徵進行分析

觀察使用決策樹的情況



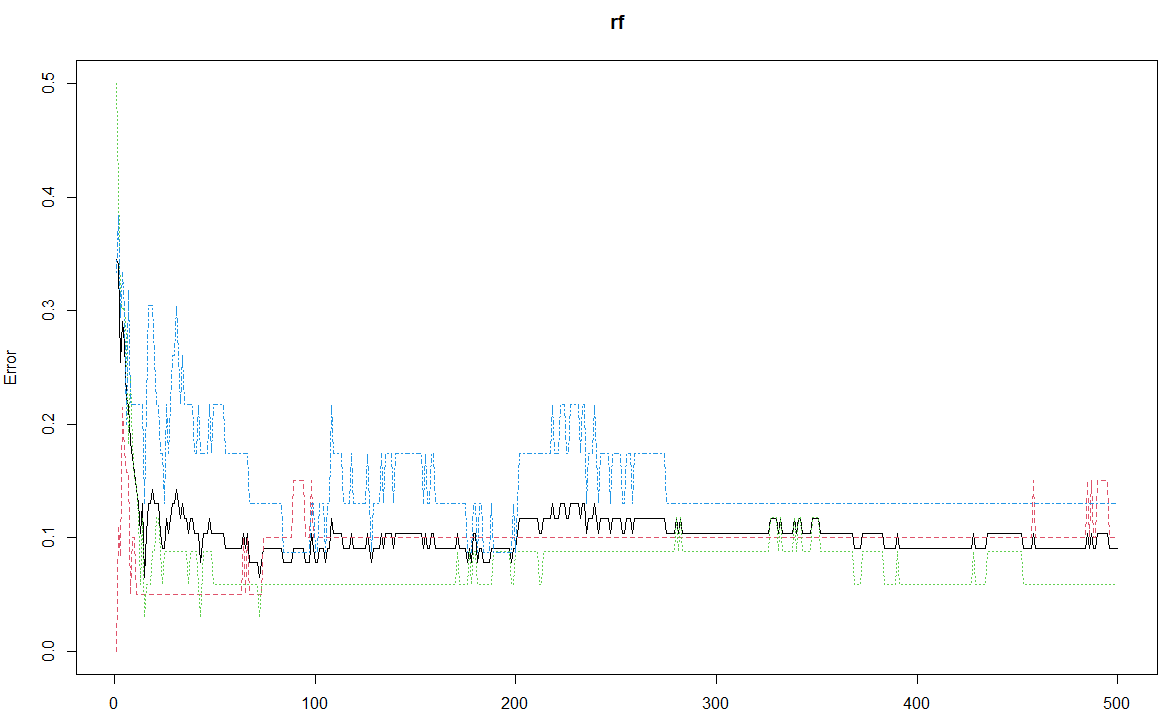


接著進入本分析重點隨機森林，這邊在R預設的隨機森林為多棵 CART樹組成，使用Gini值計算

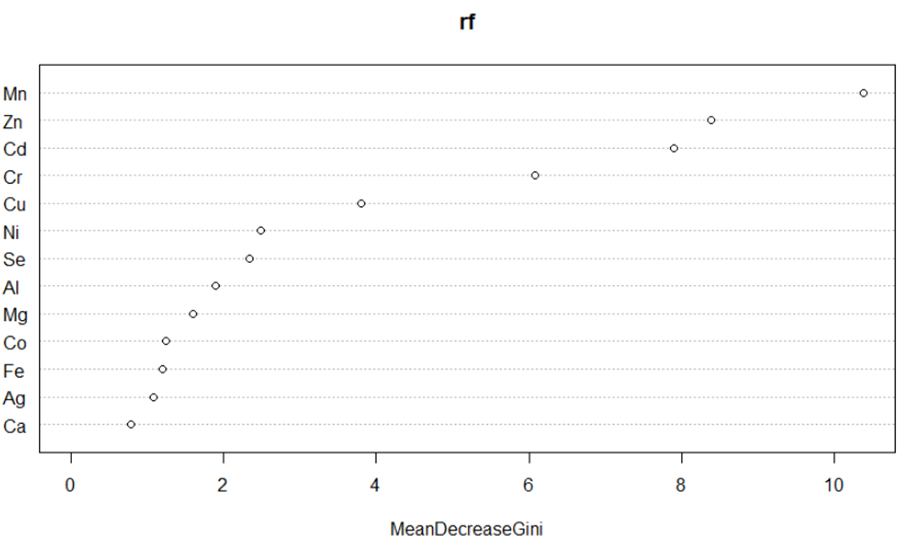
 

上圖左邊為訓練組的結果，右邊則為測試組，可以觀察到在測試組時準確率是有變好，沒有過擬情況，與決策樹的效果比較，的確隨機森林的準確率是有很大的機會是比較好。

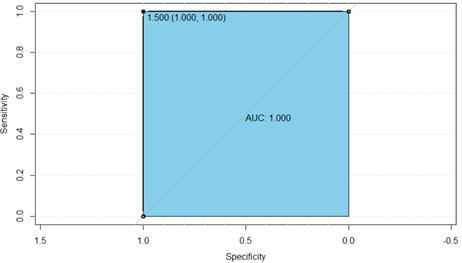
因為準確率不差，所以不對隨機森林做修剪(tuning)動作。

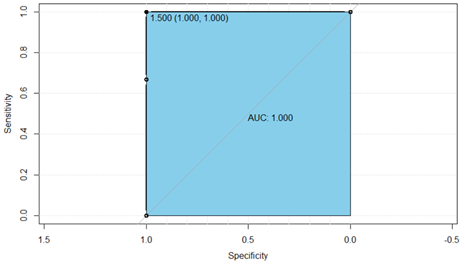


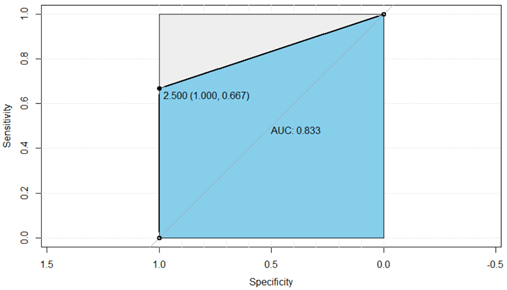
圖為error rate，綠色為class1，紅色為class2，藍色為class3，黑色為OOB。

這

變數重要性圖透過 Mean Decrease Gini 來衡量變數重要性指數，表示 Gini 係數減少的平均值。

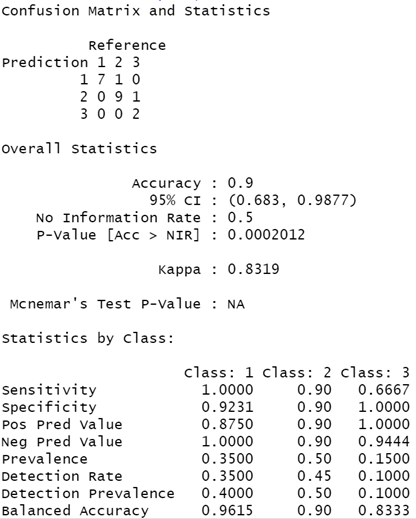


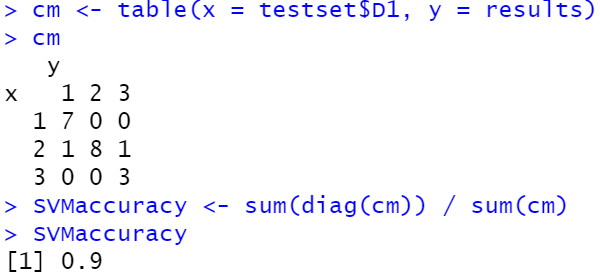


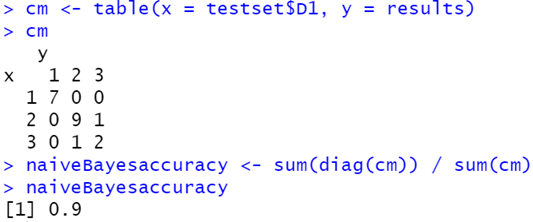


依序為class1、2、3的ROC curve，在1、2的分類情況比3還要好

最後我們將隨機森林與其他分類器進行比較，以下分別為KNN、SVM、Native Bayes的分析結果







這三種分類器的效果與隨機森林相比是差不多