**Data Mining－Project2**

**學生：吳汶峻**

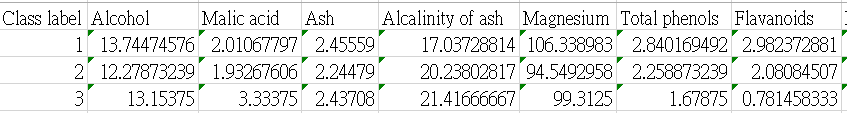
**學號：Q36071229**

**系級：電通甲**

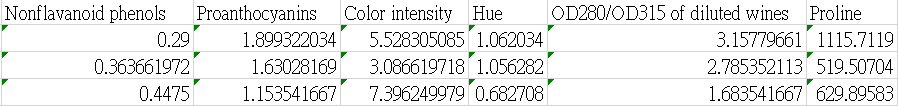
**程式與環境：Python3(Jupyter notebook)/Windows10**

1. **使用的資料集**  
   這次作業我所選擇的資料集是”Wine”，這資料集是對位於義大利的同一地區生長、但來自三種不同品種的葡萄酒進行化學分析的結果。該資料集分析確定了在這三種葡萄酒中，每種葡萄酒中都含有13種成分的濃度。其種類及意義為：  
   Alcohol：酒精  
   Malic acid：蘋果酸  
   Ash：粉煤灰  
   Alcalinity of ash：Alcalinity灰  
   Magnesium：鎂  
   Total phenols：總酚  
   Flavanoids：黃酮  
   Nonflavanoid phenols：Nonflavanoid酚類  
   Proanthocyanins：前花青素  
   Color intensity：顏色強度  
   Hue：色相  
   OD280/OD315 of diluted wines：OD280/OD315被稀釋的酒  
   Proline：脯氨酸  
     
   而實驗目的是期望能依據葡萄酒的這13種成分來預測葡萄酒的類型。  
     
   接著，我去比較Class label中每個葡萄酒品種所有的13種成份，他們各自的平均值為：

增加了Dataset Rules

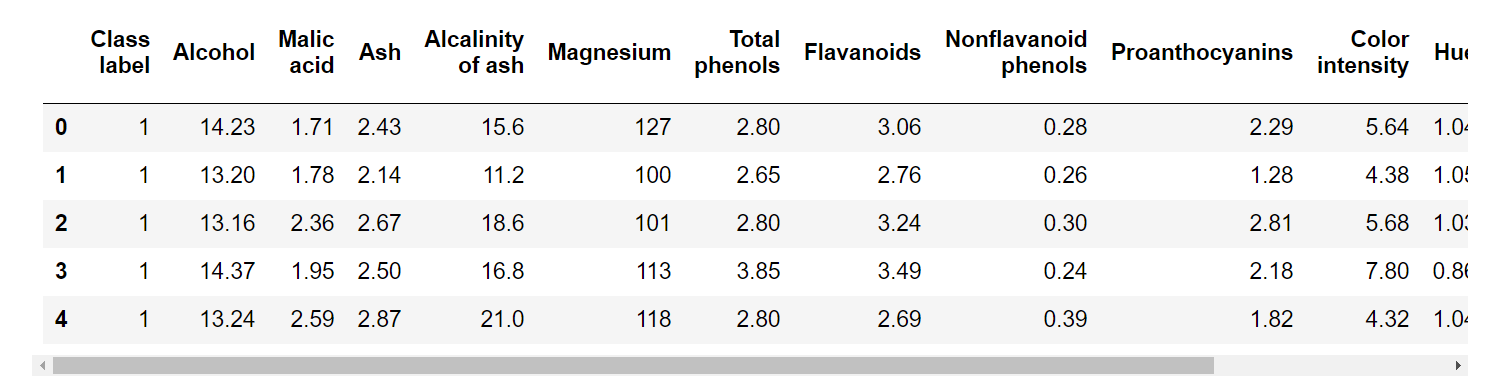
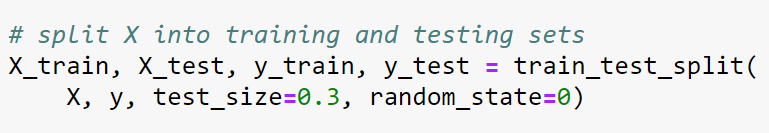
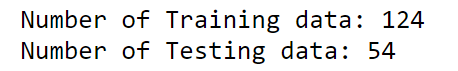
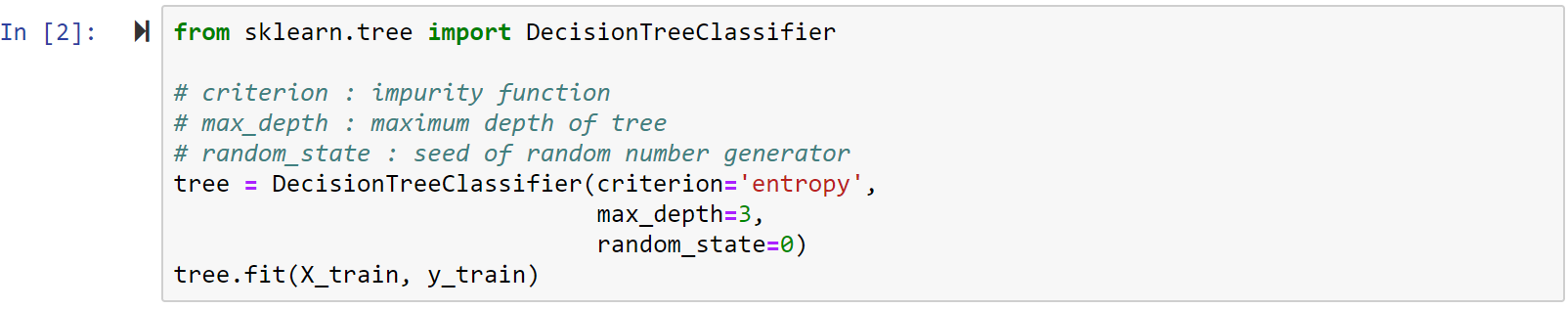
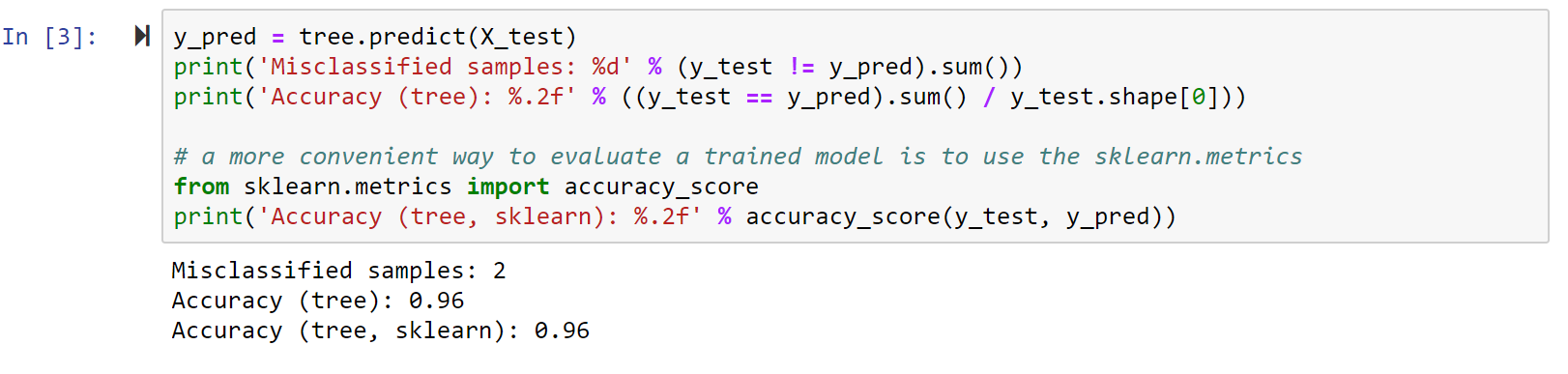


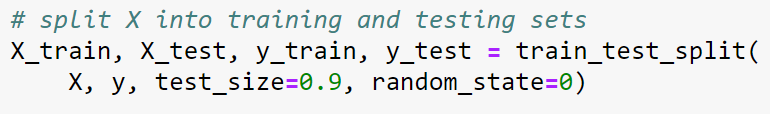
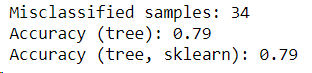
增加了Dataset Rules



從上面這兩張圖表中可以發現，從Malic acid來看，品種一與品種二的含量較品種三的低；從Magnesium來看，品種一含量普遍高於100、品種二則是低於95、品種三則是普遍在100左右；從Total phenols來看，品種一與品種二的含量平均高過2、品種三則是低於2，之後可能可以拿threshold = 2來區分出品種三；從Flavanoids來看，品種一大概平均在3左右、品種二在平均2、品種三則很明顯地低於1；從Nonflavanoid phenols來看，品種一大概在0.3左右、品種二在0.35、品種三則在0.45左右；最後的Proline，品種一幾乎都高於1100、品種二在500左右、品種三則在600左右。

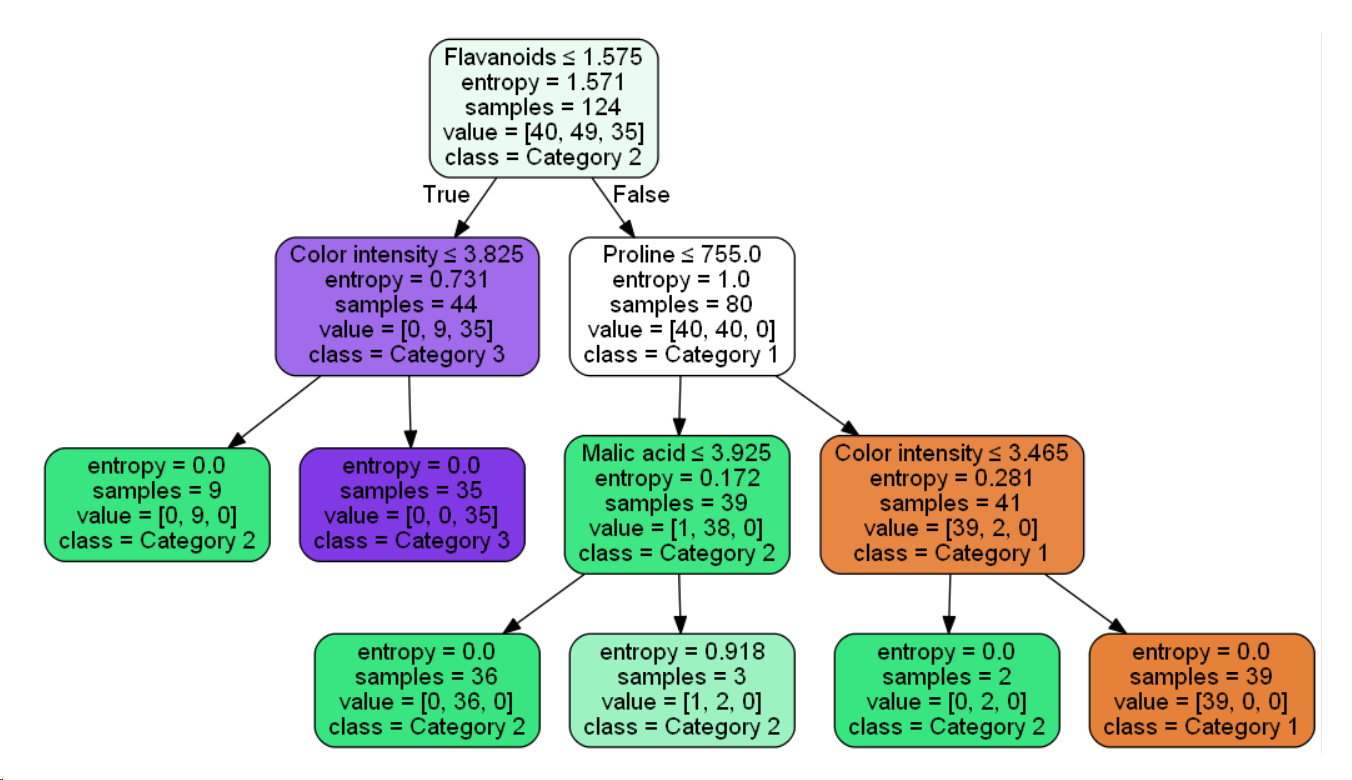
因此，從我初步的推估，若要將葡萄酒做分類，大略可以從Total phenols和Flavanoids分辨出哪瓶葡萄酒是品種三，用Proline分辨出三種不同品種的葡萄酒；而若是用Decision Tree去做區分，也可以朝這個方向去實作，但我認為每個feature之間應該也是會存在某些關係，這些都是我們要在Decision Tree裡去做探討與了解的。

1. **Decision Tree**  
   首先，我先抓資料集中前面5筆資料來檢查是否有誤，顯示結果如下：  
     
   接著，將Wine資料集隨機拆分為訓練(train)和測試(test)集，比例為7：3，程式碼及結果如下：  
     
   現在，可以利用Sklearn (scikit-learn)套件來構建Decision Tree，並檢測訓練****集與測試集的誤差值，程式碼及結果如下：

  
由上圖結果可知，在訓練集與測試集比例為7：3的情況下，此棵Decision Tree的準確度為96%，相當地高。  
  
＊補充：在經過測試後，發現到訓練集比例越低，準確度會越低。下圖為比例1：9的情況下：

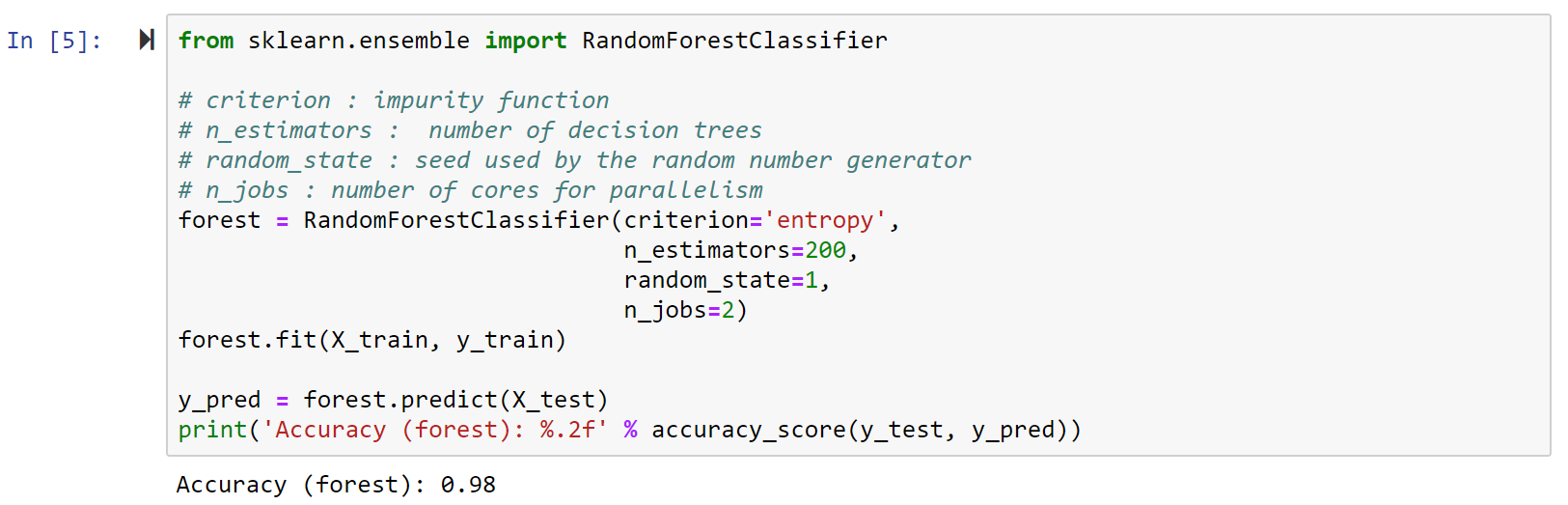
為了要讓Decision Tree更容易觀察，我們可以利用graphviz套件，讓Tree更容易視覺化。以下為程式碼：  
  


依照上圖程式碼，可畫出一棵Decision Tree。如下圖：

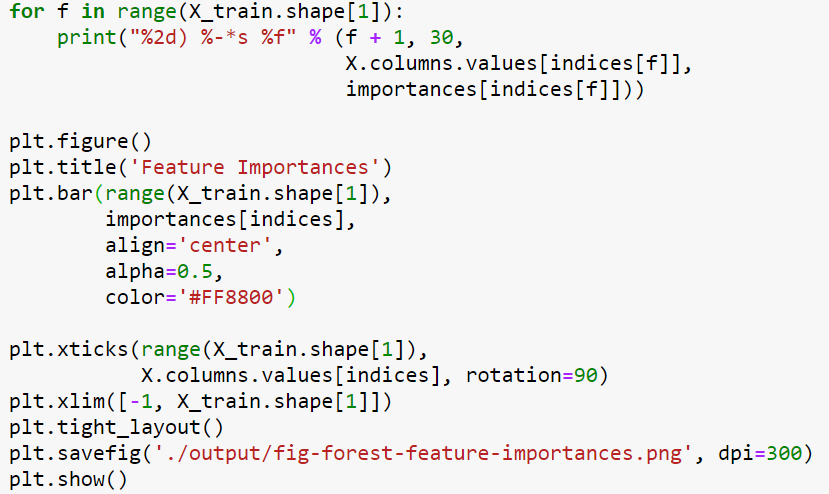
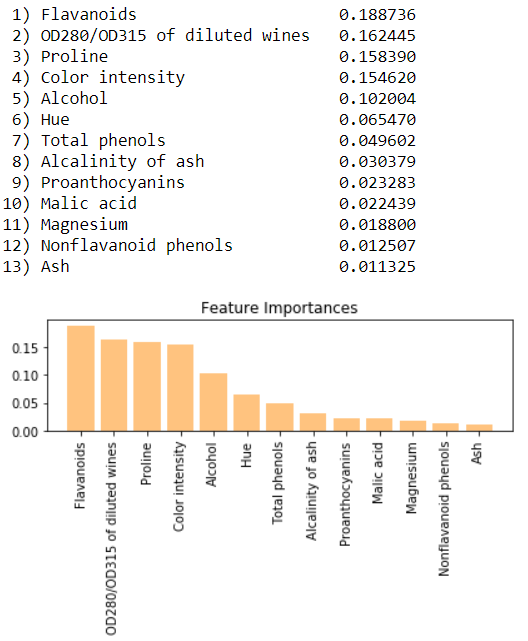
從此圖可以知道分類的依據及標準，而屬於哪一種類的葡萄酒是以框內背景顏色來做區分！

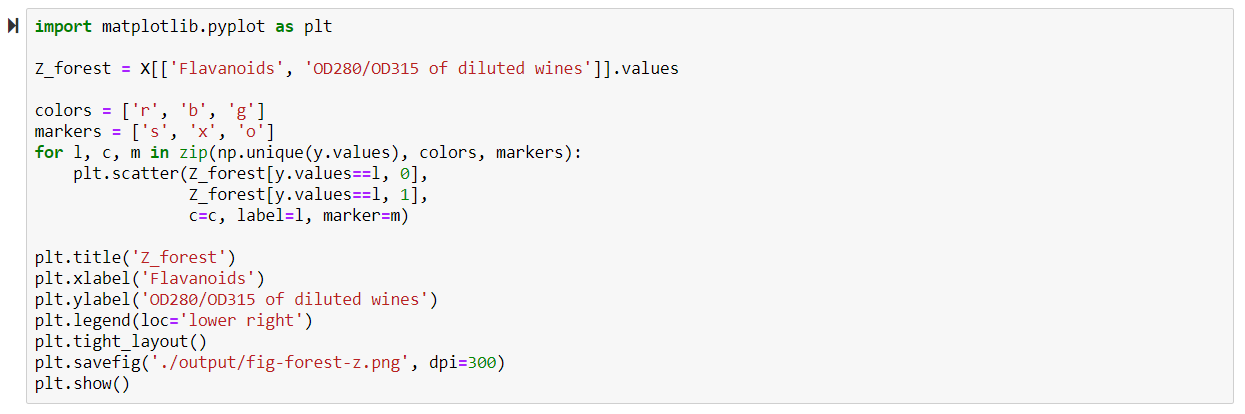
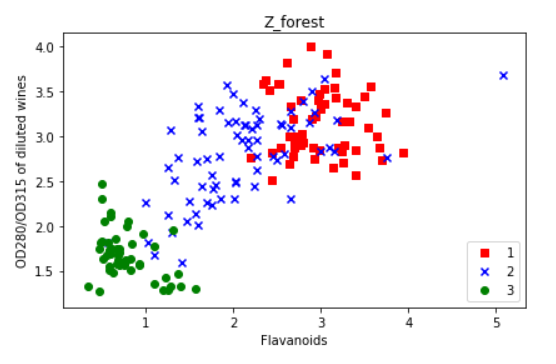
1. **Random Forest**＊概念：  
   Random Forest被廣泛運用在資料探勘中，它的效能及準確度都相當地高，而且可以觀察每一個屬性的重要度，簡單的說，Random Forest可以把它當作是多個Decision Tree組合而成的。

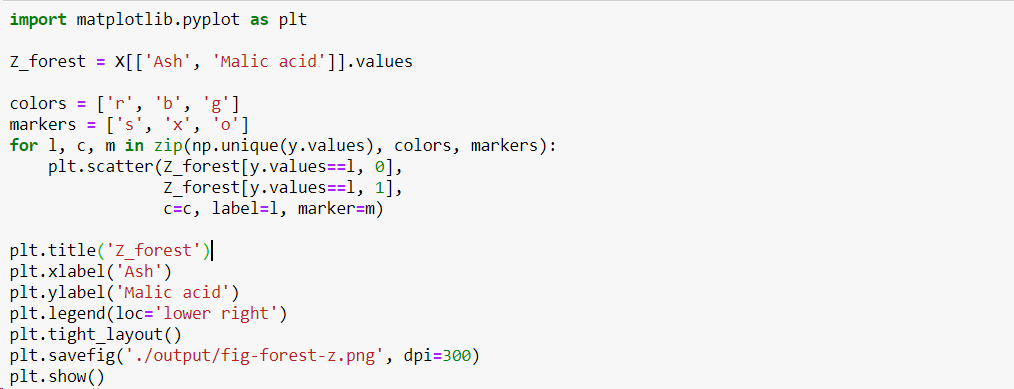
我們可以使用Sklearn的RandomForestClassifier套件去創建一個Random Forest，程式碼及結果顯示如下：

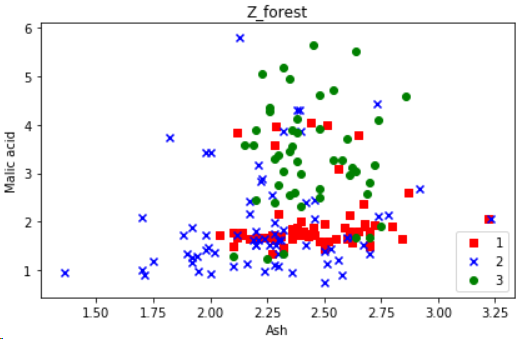


由此圖可知Random Forest的準確度為98%，比Decision Tree準確要更高、預測得更準！

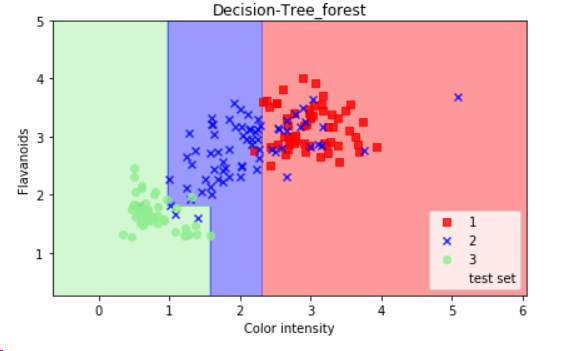
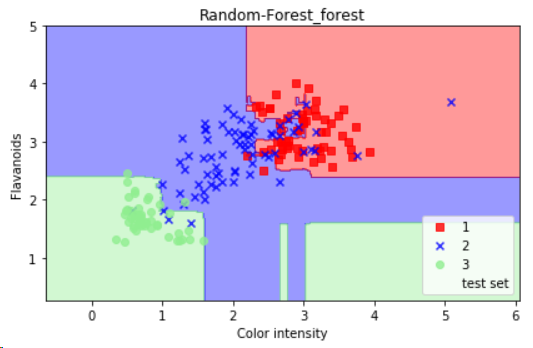
此外，可以利用Random Forest來計算屬性重要度，這樣可以將比較不重要的屬性刪減，可以讓預測與準確度更高。程式碼及結果如下顯示：

在刪減一些不重要的屬性後，得到了Flavanoids與OD280/OD315 of diluted wines是這個資料集重要度最高的，將此兩項屬性萃取出來並作圖，結果可以蠻容易區分葡萄酒種類：

  
我們可以做測試，若將某兩個對資料集較不重要的兩個屬性，如：Ash與Malic acid拿來作圖，會產生葡萄酒種類無法區分的結果：



總結：重要度低的屬性根本區分不了葡萄酒的種類，因此取重要度高的屬性也是相當重要的一環。也可知道，並不需要全部屬性去偵測，只需要將重要度高的屬性抓出來比對，就能大略分類出葡萄酒的種類。

另外，為了比較Decision tree及Random Forest精準度，可以畫出兩種的決策邊界，下圖為Decision-tree與Random-Forest的決策邊界：

由這兩張圖做比較後，可以發現Random Forest比Decision Tree更為精準些！

1. **實作心得**這份作業需要利用Decision Tree來估測資料集的分類。剛開始還不太懂Decision Tree背後原理以及如何將資料做分類，但在上網看很多有關Decision Tree 程式的邏輯及演算法方向後，從怎麼樣找到好的資料集、做資料處理、到用程式畫出Decision Tree最後再做分析及比對預測與實際結果，可知道用Decision Tree做資料探勘其實是件不簡單的事，但只要肯學，其實是可以融會貫通的。而且也藉由此次機會，不只學到Decision Tree分類法，也學到Random Forest的應用，希望未來在需要探勘資料時，能夠將今日所學學以致用！