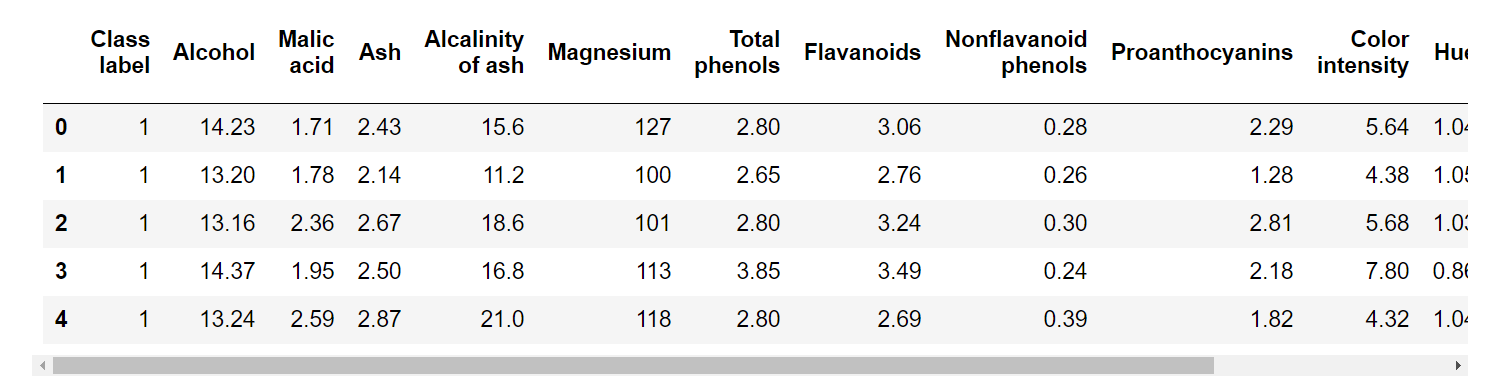
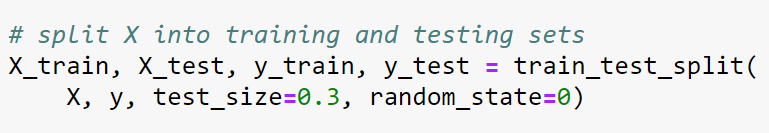
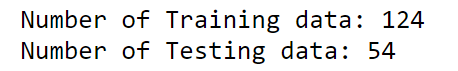
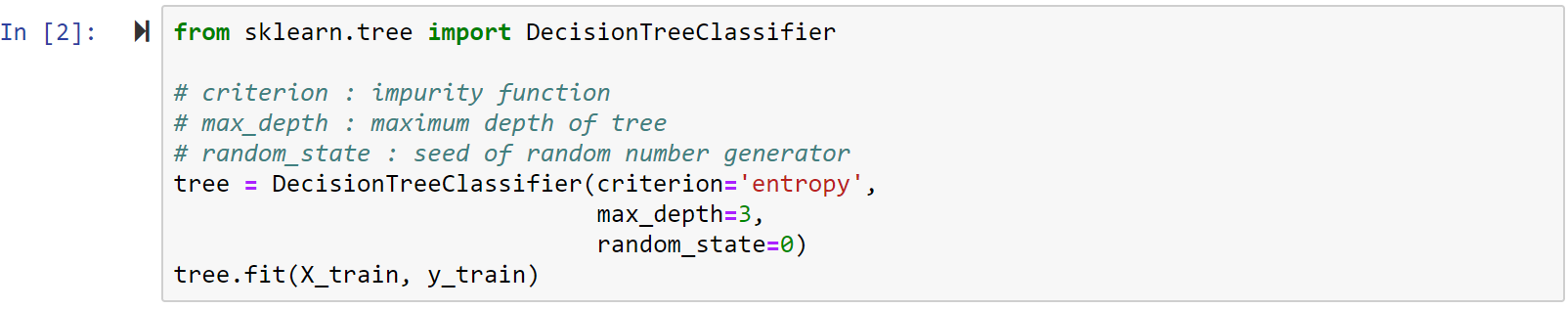
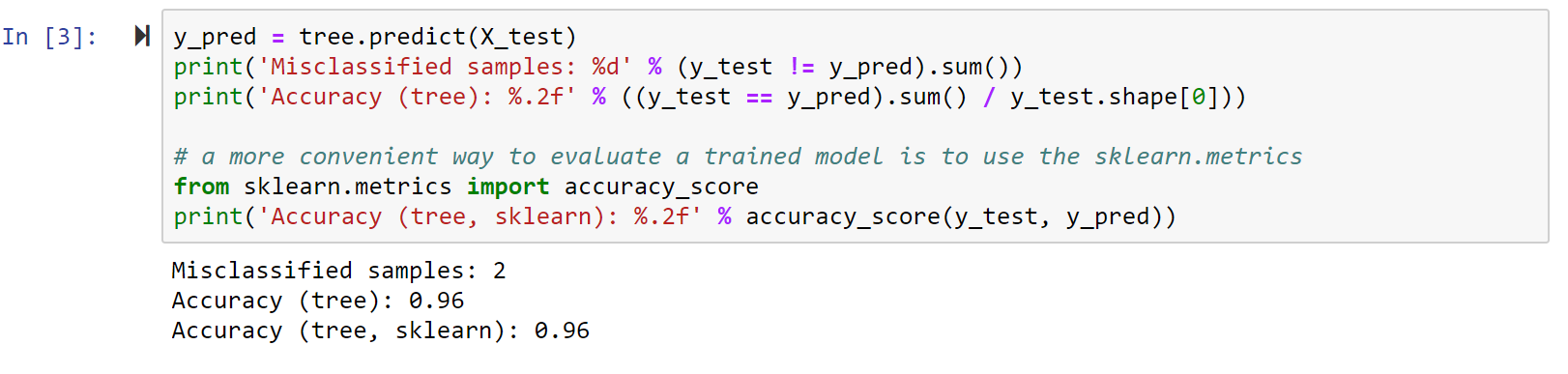
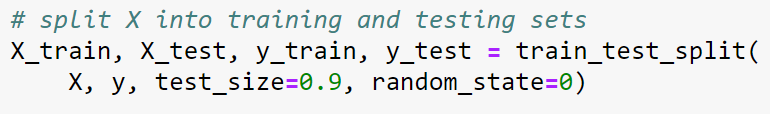
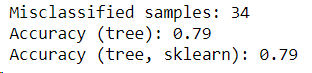
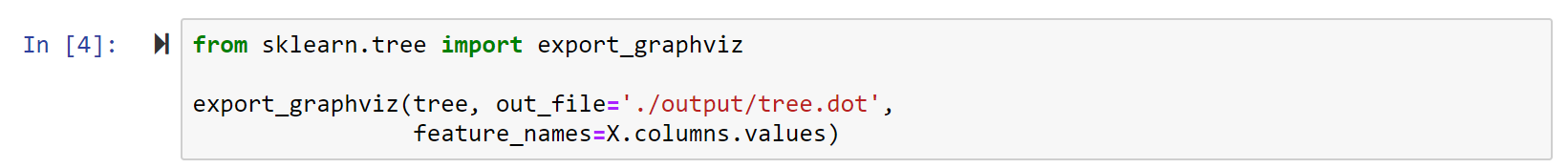
**Data Mining－Project2**

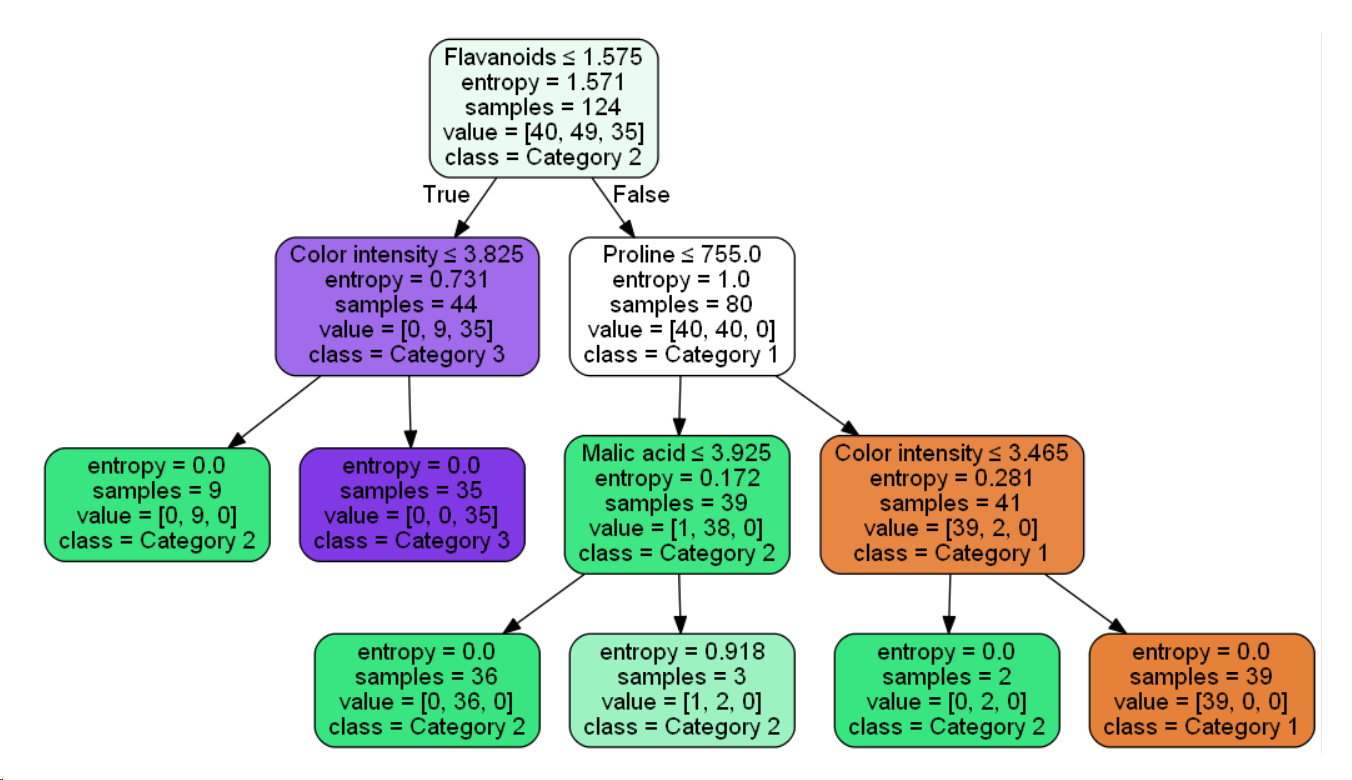
**學生：吳汶峻**

**學號：Q36071229**

**系級：電通甲**

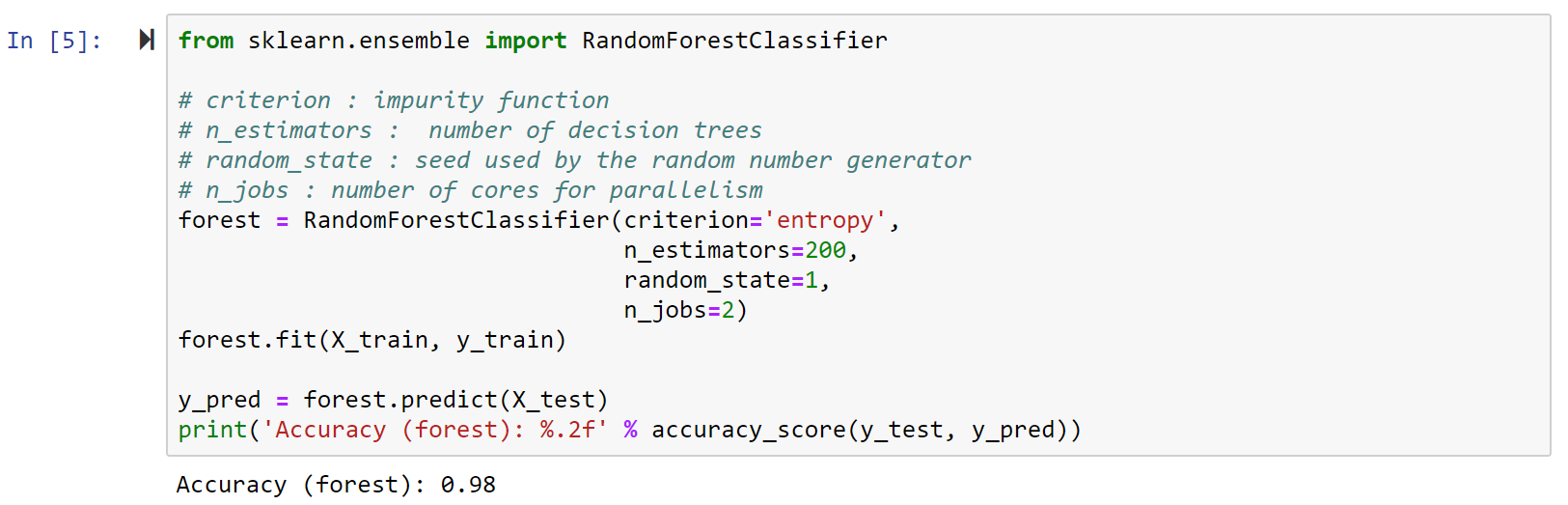
**程式與環境：Python3(Jupyter notebook)/Windows10**

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。
2. **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%，相當的高。  
     
     
     
   ＊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：  
     
     
     
   此時會產生一個.dot檔，在所在資料夾內執行指令：「dot -Tpng tree.dot -o fig-tree.png」，就可以畫出一棵Decision Tree。如下圖：

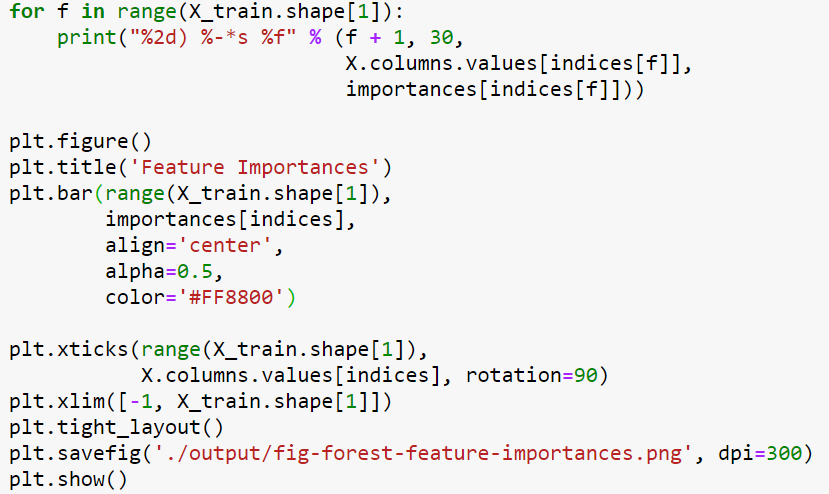
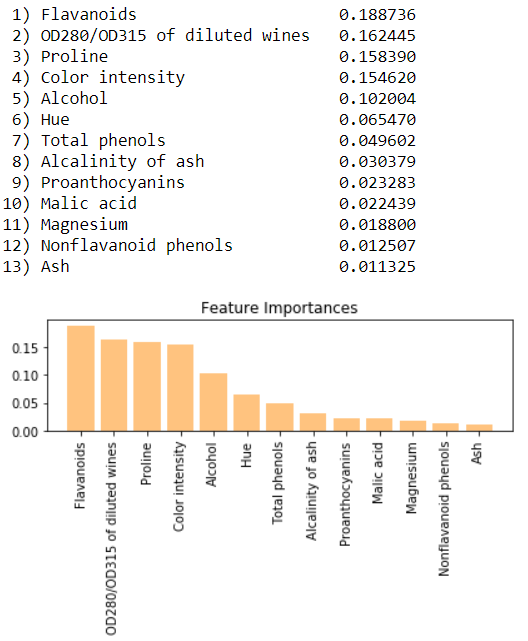
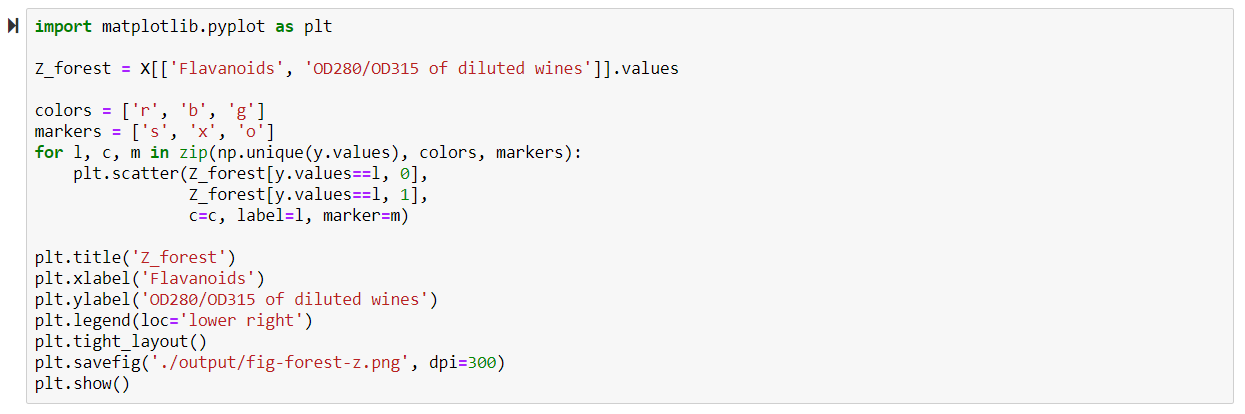
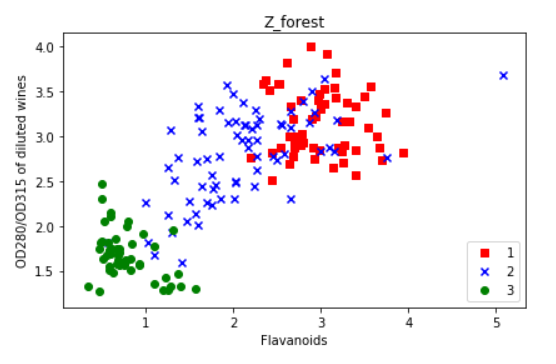
從此圖可以知道，Flavanoids<=1.575時，可以將第一種與第三種類的葡萄酒做分類，十分有效率！

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：

並不需要全部屬性去偵測，由此圖可知，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。

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到底要怎麼做搜索
2. …
3. …
4. …