資料探勘 HW01 科管所一 0753524 邱薇如

1. **把資料讀取進來(可用pandas套件)**

|  |
| --- |
| data = pd.read\_csv('D://character-deaths.csv')  data |
|  |

1. **資料前處理**

**2.1 把空值以0替代**

|  |
| --- |
| data = data.fillna(0)  data |
|  |

* 1. **Death Year , Book of Death , Death Chapter三者取一個，將有數值的轉成1**

選擇 Death Chapter 欄位作為 Label ， 找到 Actual Data，並命名為Y

|  |
| --- |
| Y = data['Death Chapter']  Y = pd.DataFrame(np.where(Y>0,1,0))  Y.columns=['Actual Data']  Y |
|  |

* 1. **將 Allegiances 轉成 dummy 特徵**

將變數項移除，並將 Allegiances 轉換成 dummy 特徵

|  |
| --- |
| X = data.drop(['Name','Death Year','Book of Death','Death Chapter'],axis=1)  X = pd.get\_dummies(X, columns=['Allegiances'])  print(chd['Allegiances'].value\_counts()) |
|  |

**2.4 亂數拆成訓練集(75%)與測試集(25%)**

|  |
| --- |
| from sklearn.model\_selection import train\_test\_split  X\_train,X\_test,Y\_train,Y\_test = train\_test\_split( X,Y,test\_size=0.25,random\_state=345723 ) |
|  |

1. **使用 scikit-learn 的 DecisionTreeClassifier 進行預測**

決策樹預測準確度為0.76419

|  |
| --- |
| from sklearn import tree  clf = tree.DecisionTreeClassifier(max\_depth=5)  clf = clf.fit(X\_train,Y\_train)  pred = clf.predict(X\_test)  print('Accuracy of training set',clf.score(X\_train,Y\_train)) |
|  |

1. **做出Confusion Matrix，並計算Precision, Recall, Accuracy**

Confusion Matrix 成功計算出

Accuracy：0.70434 / Precision：0.55421 / Recall：0.59740

|  |
| --- |
| CF\_M = pd.DataFrame( confusion\_matrix(Y\_test, pred),  index=['Actual Alive','Actual Death'], columns=['Predicted Alive','Predicted Death'])  print('Testing Set')  print('Accuracy:',accuracy\_score(Y\_test, pred))  print('Precision:',precision\_score(Y\_test, pred))  print('Recall:',recall\_score(Y\_test, pred)) |
|  |

|  |
| --- |
| import graphviz  dot\_data = tree.export\_graphviz( clf,out\_file=None, feature\_names=X\_test.columns, class\_names=[‘Alive’,’Death’], max\_death = 5)  graph = graphviz.Source(dot\_data)  graph |

1. **產出決策樹的圖**

