科目：1071_資料探勘 DATA MINING

學生：楊沛霖

學號：Q36071156

日期：2018-11-18

作業要求：

1. 本次作業是自己設計 DataSet (包含特徵值) ,並且要求作業要有正確的規則來生成。
2. 除了生成上述的 DataSet 還要產生 Decision tree 或是其他驗證模型，驗證自己的 DataSet。
3. 討論 Decision tree 的規則與一開始設計的規則的關係。

環境與程式語言：

資料生成：excel

程式環境：jupyter 上利用 ipython 運作

一、 資料生成

Department and curriculum closeness rate (系所與課程的相近度) :

由常態分布 mean = 70 , σ = 15 生成

Attendance rate (出勤率) :

由常態分布 mean = 70 , σ = 15 生成

HW score (作業成績) :

由常態分布 mean = 80 , σ = 10 生成

Report score (報告成績) :

由常態分布 mean = 80 , σ = 15 生成

number of ask teacher or TA (詢問老師或 TA 的次數) :

由常態分布 mean = 2 , σ = 2 生成

Test score (考試成績) :

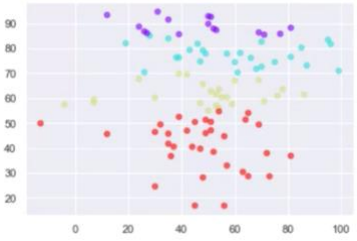
此為我們做 Decision tree 的區分標準, 由上述資料依比例計算出來,級距為 15 而區間範圍為

級等	Range
A	85-100
B	71-84
C	56-70
D	40-55
E	40以下

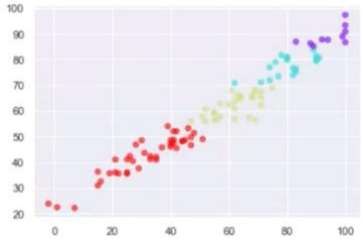
二、 資料分析

本次的生成 100 筆資料而這次 Test score 生成方式是讓 Attendance rate (出勤率)佔所有項目中最高的 0.7(其他特徵值接不超過 2)，因此預估在關聯度上會看到 Test score 和 Attendance rate 有很大的關聯性。

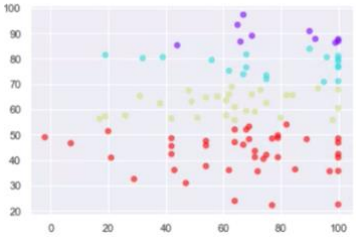
下圖為 Y 軸均為 Test score 而 X 軸分別為：系所與課程的相近度, 出勤率, 作業成績, 報告成績, 詢問老師或 TA 的次數



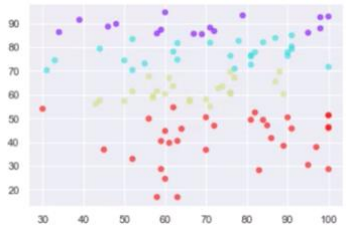
圖一、系所與課程的相近度.



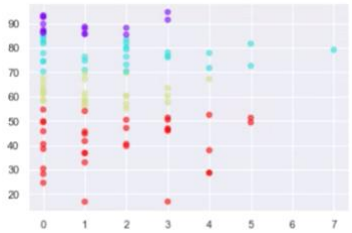
圖二、出勤率



圖三、作業成績



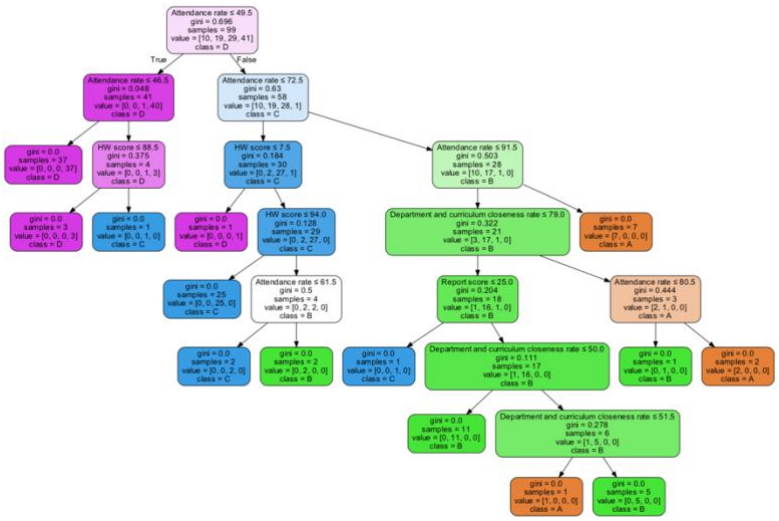
圖四、報告成績



圖五、詢問老師或 TA 的次數

由圖片可以些明顯看出出勤率與分數有非常大的正相關，而其他資料則是非常散狀的隨機排列，尤其是 詢問老師或 TA 的次數 ，因此項是所以特徵值裡比重佔最小。

Decision tree 則為(顏色為不同的分數區間)：



圖六、用 Test score 來區分的 Decision tree

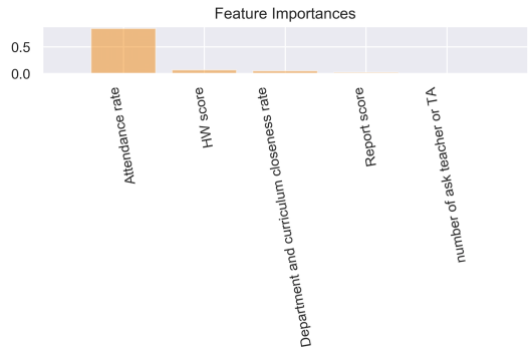
由上圖的 Decision tree 中可以看到在 13 個分岔中光是 Attendance rate 就佔了 6 個分岔了也表示此特徵值是很重要的決策根據。

三、 Decision tree 驗證

這次用 Decision tree 所跑出來的分類結果在與額外生成的測試資料實際丟下去運行後的結果有 75.5%正確，還算可以。原先原本想用更多筆資料下去跑 model，但是發現訓練資料數量越多 Decision tree 的分支也就越多，因此訓練資料不敢放太多資料下去，不然自己生成的資料當然是要多少有多少。

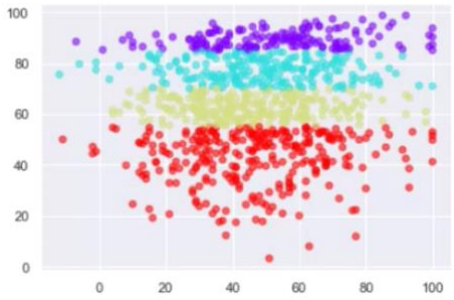
此為用 DecisionTreeClassifier 裡的 feature_importances_ 所找出來的特徵值重要度，可以看到其配給的比例與自己設計的配分比例有一定的相似度(照順序為 0.7、0.125、0.1、0.025)

```
1) Attendance rate                0.848089
2) HW score                       0.073086
3) Department and curriculum closeness rate 0.052916
4) Report score                   0.025909
5) number of ask teacher or TA    0.000000
```

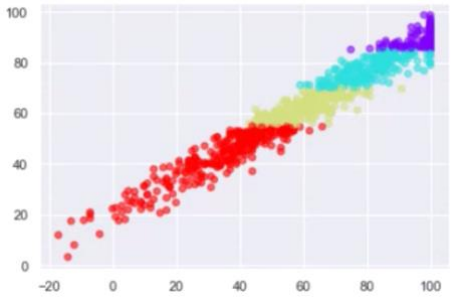


四、 大量資料(900 筆)下的結果與驗證(100 筆)

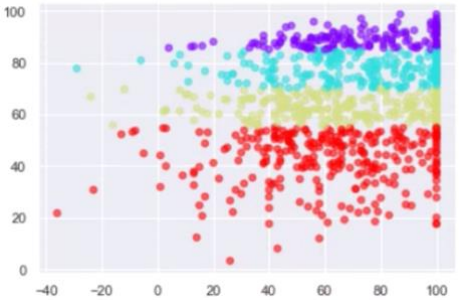
下圖為 Y 軸均為 Test score 而 X 軸分別為 系所與課程的相近度, 出勤率, 作業成績, 報告成績, 詢問老師或 TA 的次數



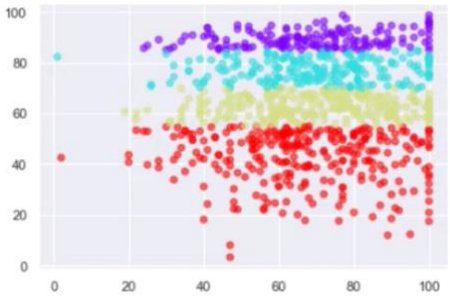
系所與課程的相近度



出勤率



作業成績



報告成績

相比少量資料，在大量資料下更可以明顯地看出 系所與課程的相近度 比起其他三項都還要沒有那麼強的關聯度，而作業成績又比報告成績略高了點。

而這次的特徵值重要度又更貼近實際給予的比重了(剛剛為 0.84、0.073、0.053、0.026、0)。 (實際給予的比重照順序為 0.7、0.125、0.1、0.025)

而利用模擬跑出來的決策樹準確率也有所提升(75.5%➔89%)。

```
1) Attendance rate                        0.686001
2) HW score                               0.171876
3) Department and curriculum closeness rate 0.080165
4) Report score                           0.055674
5) number of ask teacher or TA            0.006284
```

五、 心得

這次最難處理的部份是在資料的生成，原本的那些項目都還會依照實際情況給予比較好看的數據，例如作業成績都 80up 或是報告成績都不差等等的，但是這些資料跑出來的圖形就變得很不好分辨，除非利用比較極端的數值不然真的很難一下子馬上就看出每個圖之間的差別。

而數據的多寡也是一大難題，當數據太多就會出現很複雜的決策樹，複雜到根本沒辦法閱讀...