國立成功大學

資料探勘 Data Mining

Project02

課程教授：高宏宇

學生：葉芯妤

學號: P96074147

目錄

一、目標說明 ............................................................... 3

二、資料說明 ............................................................... 3

三、實作說明 ............................................................... 5

四、分析比較 ............................................................... 7

1. 目標說明

2

3

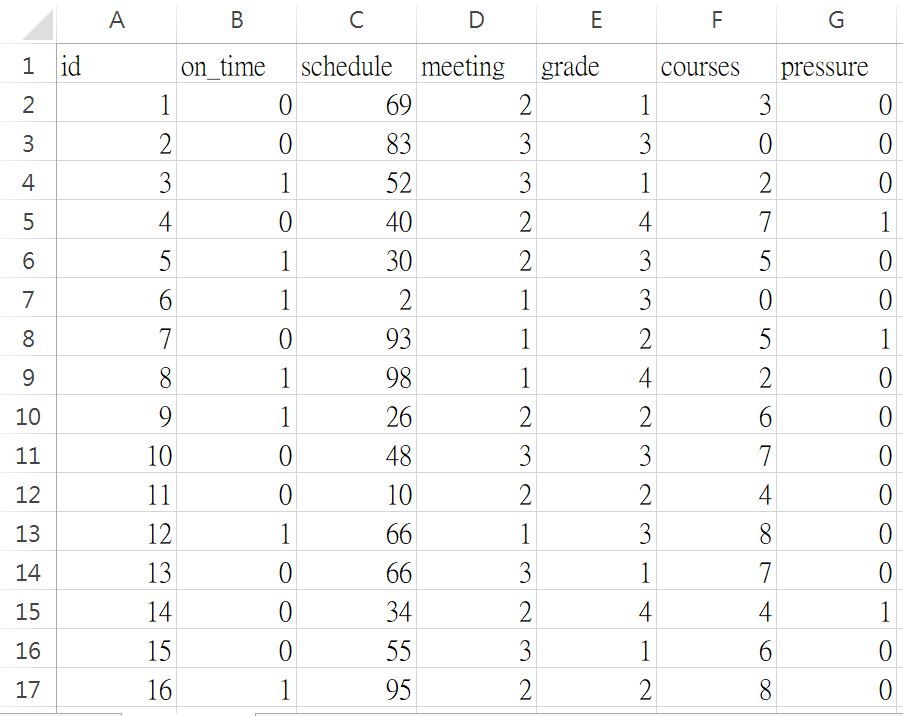
1

二、資料說明

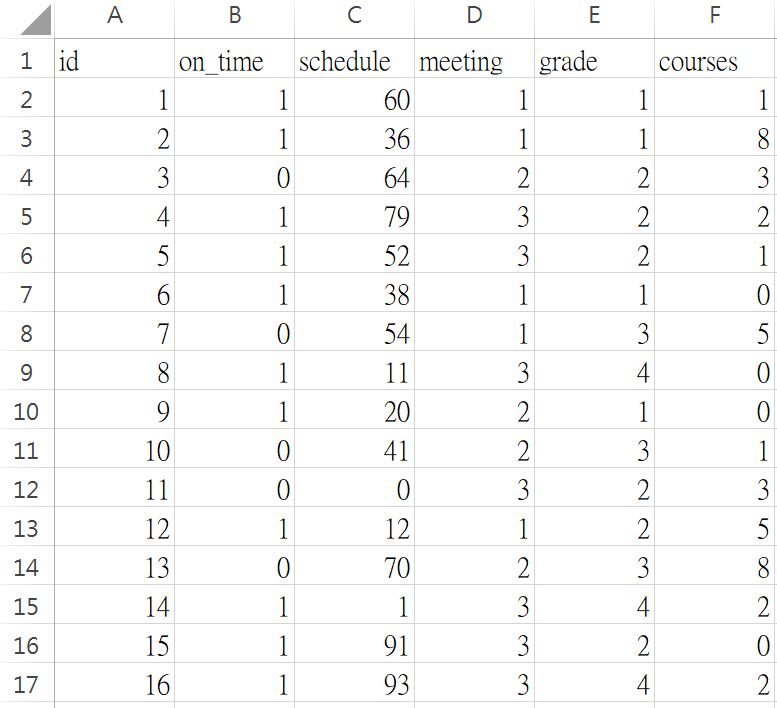
以預測研究生的壓力為主軸，建立資料，內容包含是否能準時畢業、論文進度、開會頻率、年級以及修課數量，下表為欄位說明。共有5個分類特徵，黃色部分為分類的項目，即為研究生有無壓力。

|  |  |  |
| --- | --- | --- |
| 欄位 | 說明 | 內容 |
| on\_time | 是否能準時畢業 | 是=1，否=0 |
| schedule | 論文進度 | 未完成一半=0~49，完成一半=50~100 |
| meeting | 開會頻率 | 低=1，中=2，高=3 |
| grade | 年級 | 一年級=1，二年級=2，三年級=3，四年級=4 |
| courses | 修課數量 | 小於等於四門課=0~4，大於四門課=5~8 |
| pressure | 研究生壓力 | 有=1，無=0 |

利用excel建立資料，並用亂數產生100筆資料，命名為data\_train.csv，下圖為資料內容：



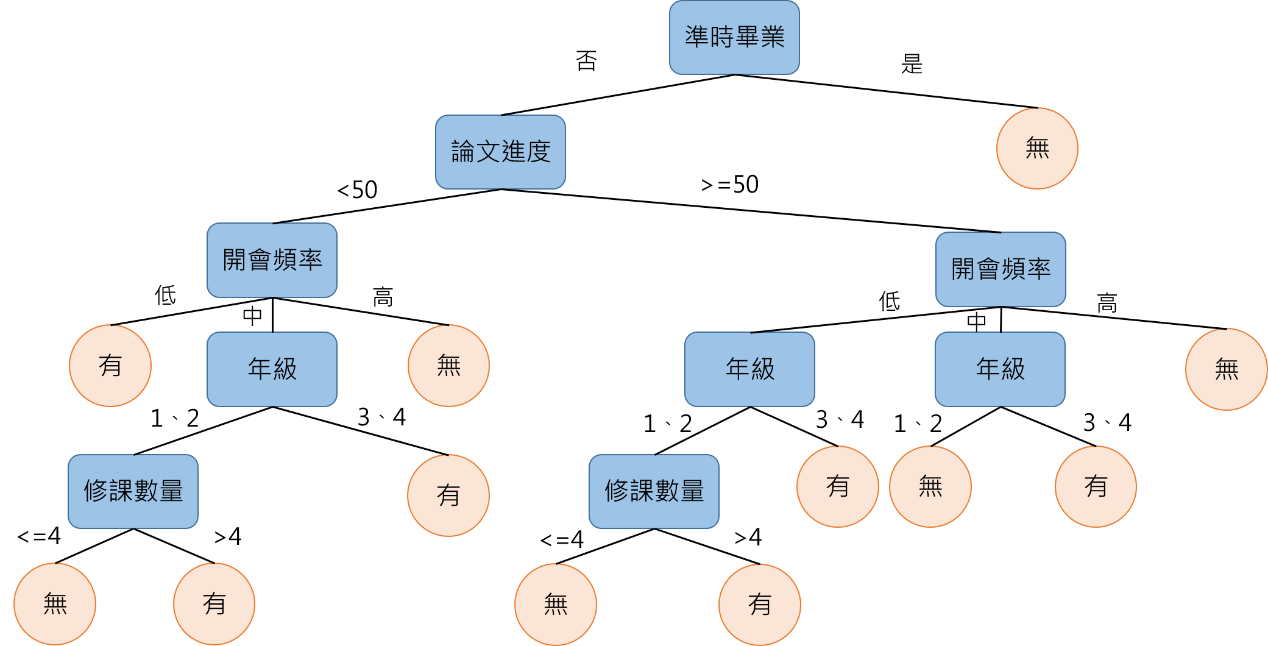
另外建立一測試的資料，同樣利用excel建立，並用亂數產生100筆資料，命名為data\_test.csv，下圖為資料內容：



* 我的預測：

從資料中觀察到，是否能準時畢業為研究生最大的壓力來源，因此，我以是否準時畢業作為考量，認為此關係影響較大，畫出決策樹。下圖為預測的決策樹，規則為：

1. 研究生是否能準時畢業
2. 論文進度是否完成一半
3. 開會頻率多寡
4. 研究生的年級
5. 修課數量是否大於四門課



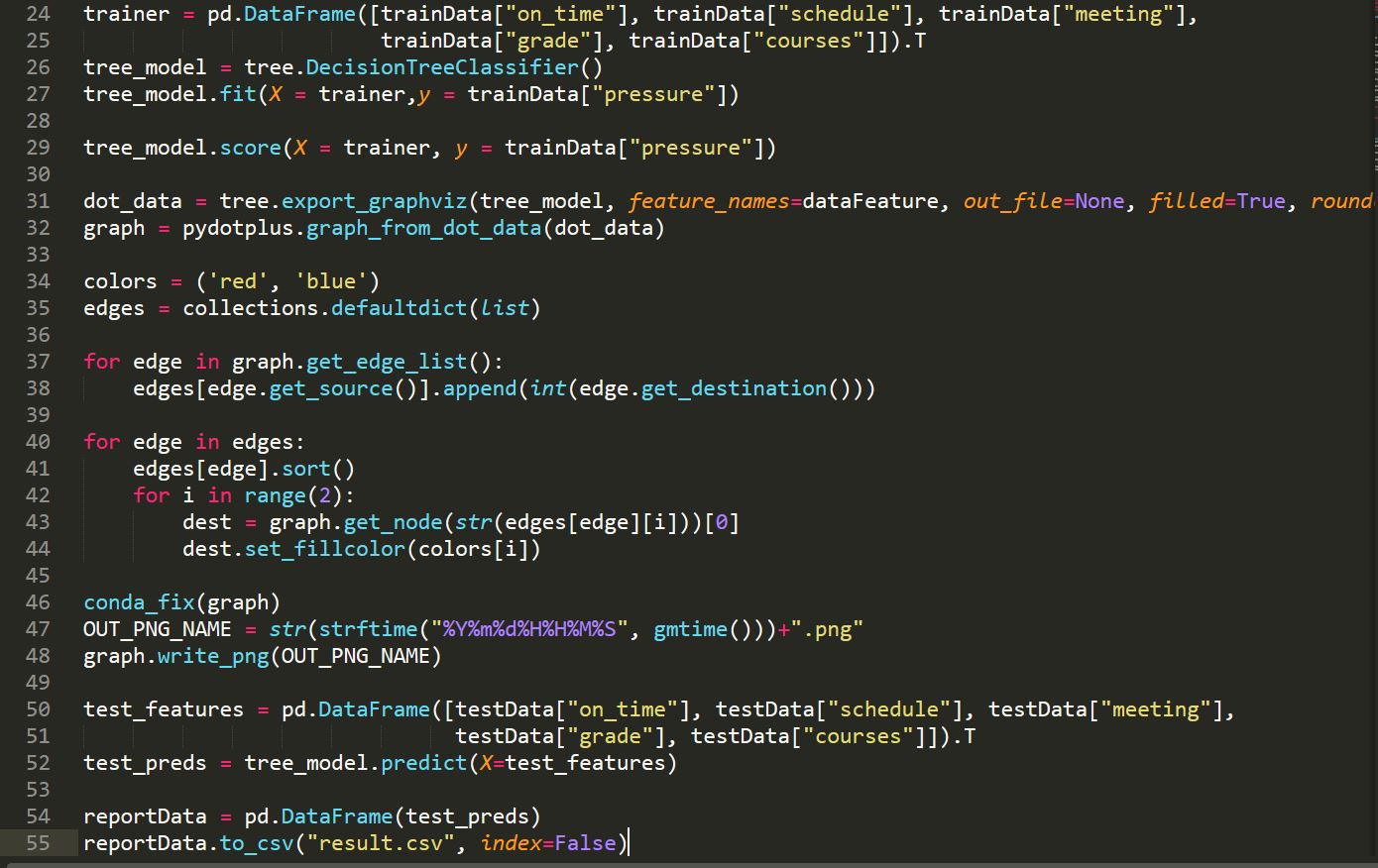
三、實作說明

利用Python撰寫程式碼，步驟為：先將檔案做前處理後，再訓練決策樹，將決策樹視覺化輸出，最後輸出預測模型，下圖為實作步驟及程式碼。

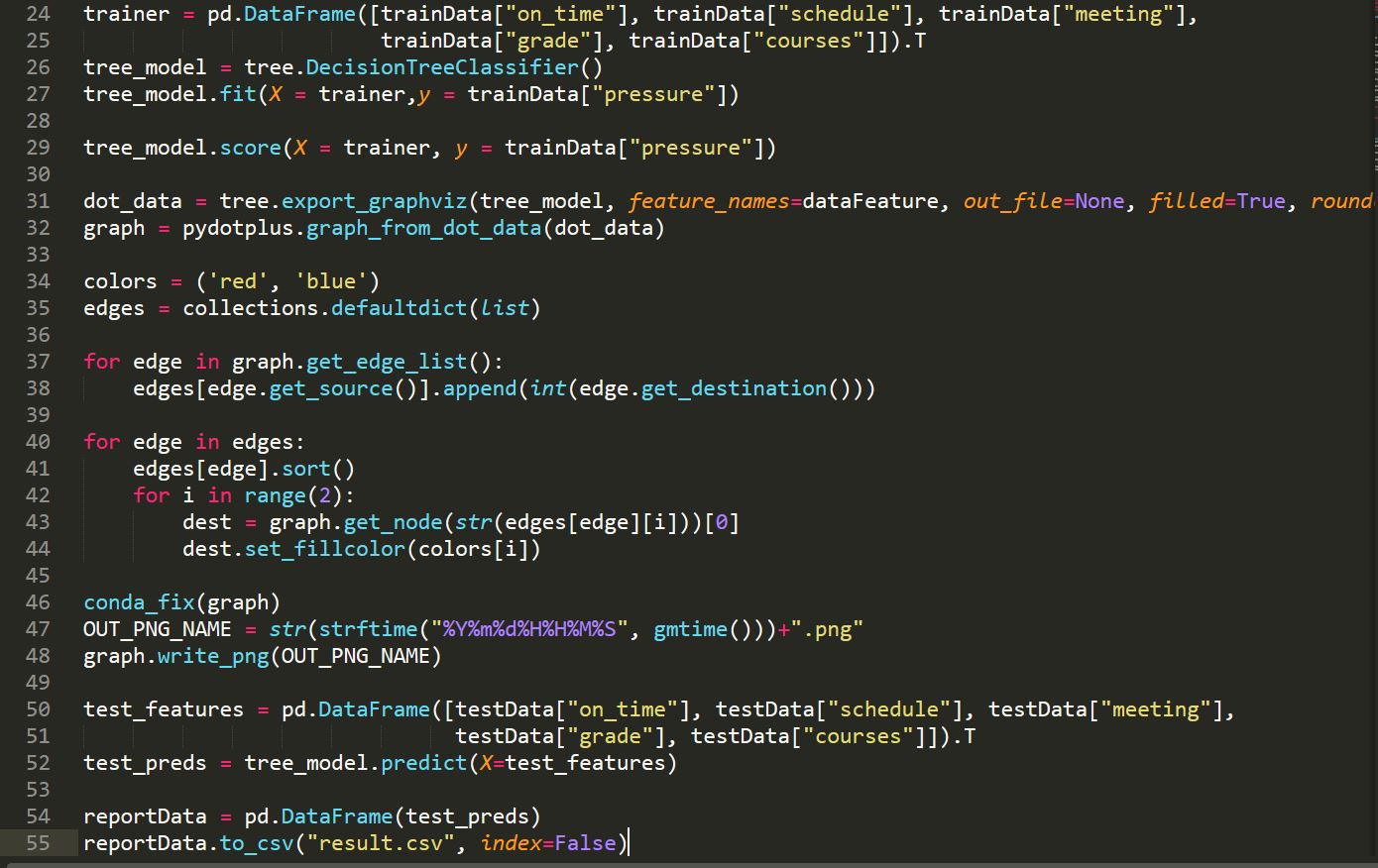
1. 檔案做前處理



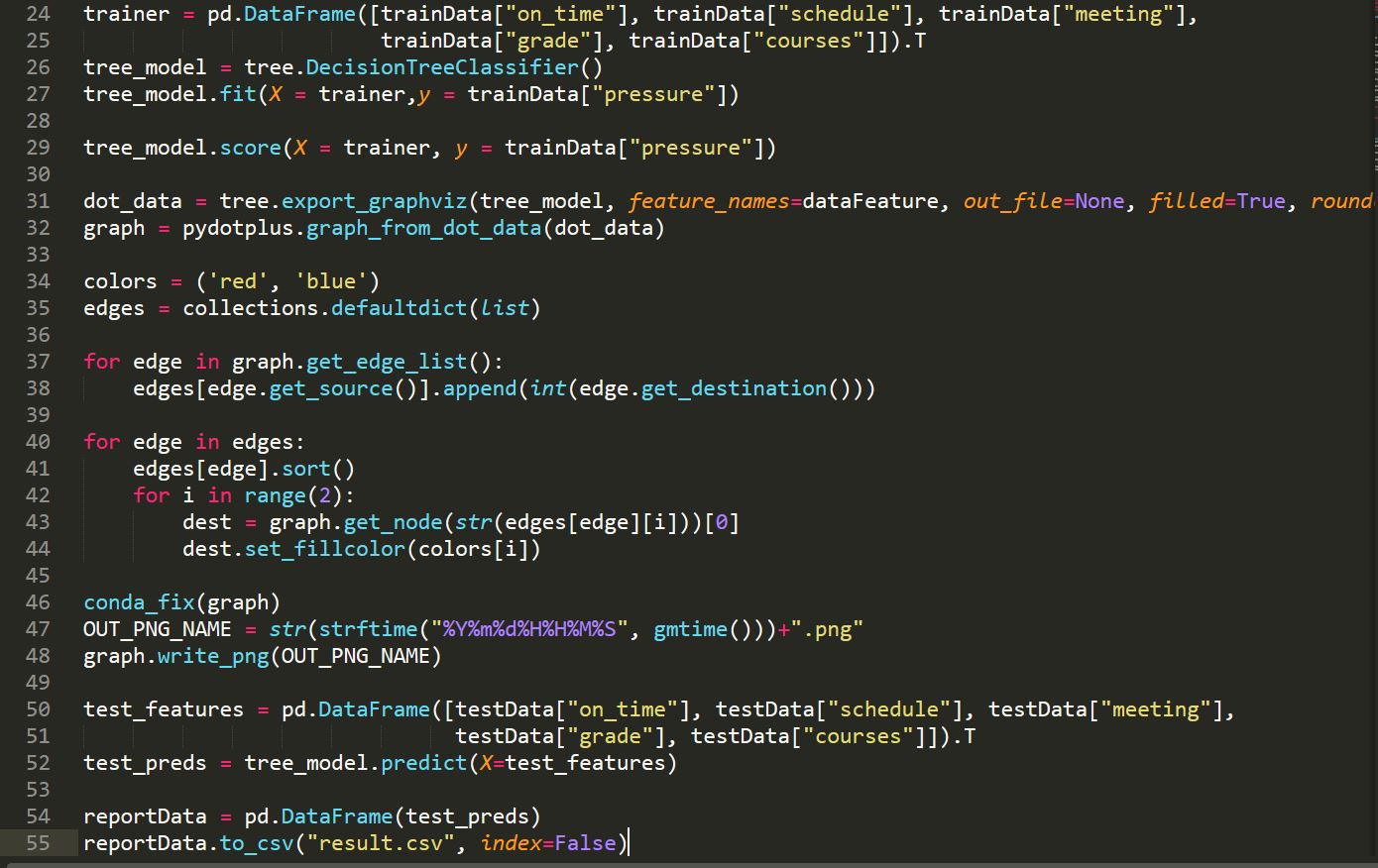
1. 訓練決策樹



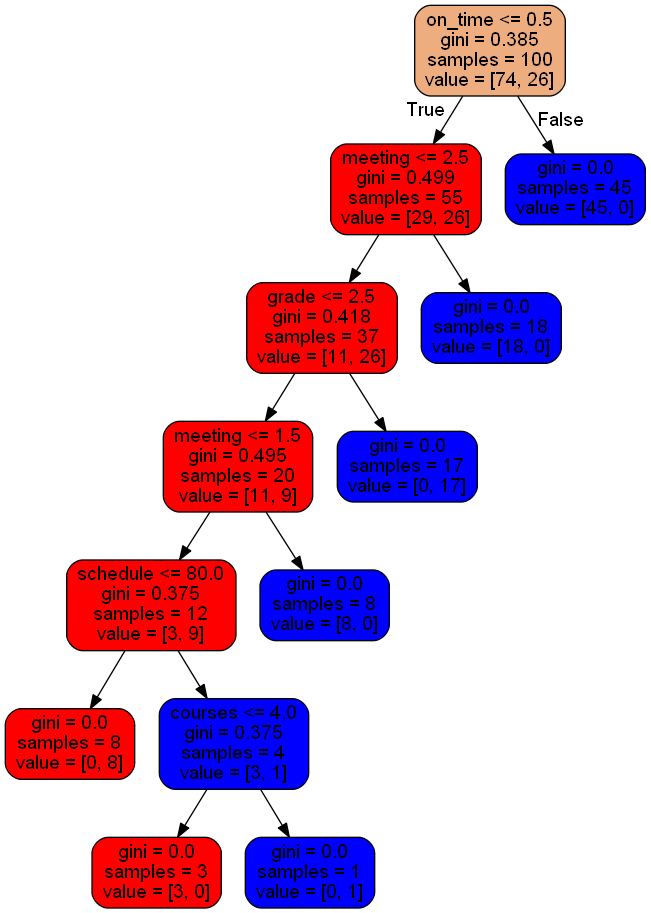
1. 決策樹視覺化輸出



1. 輸出預測模型



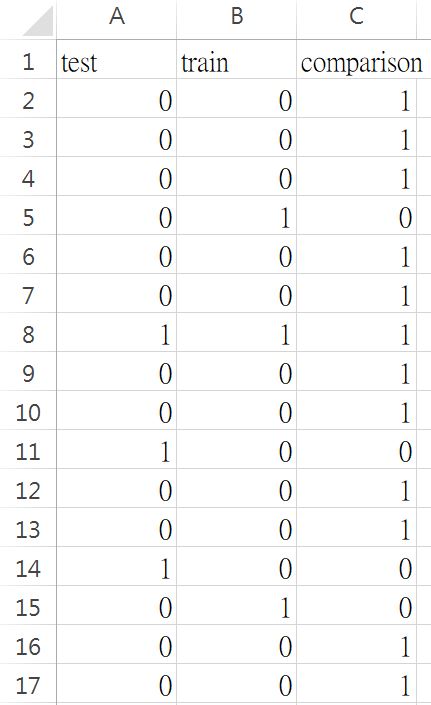
最後輸出的結果命名為result.csv，將針對產生結果與原先訓練的資料進行比對，而產生的決策樹如下圖所示：



四、分析比較

將data\_test.csv的訓練結果與原本data\_train.csv的結果進行比對，另外建立一excel檔，命名為comparison.csv，從全部100筆資料中取四分之一出來進行分析，也就是25筆資料。

比對結果如下圖所示，欄位test為data\_test.csv的訓練結果，而欄位train為原本data\_train.csv產生的結果，若訓練結果與實際結果是一樣的，即為0和0或1和1，那麼比較結果，也就是欄位comparison，則為1；反之則為0，代表結果為0和1或1和0。



觀察結果發現25筆資料中有7筆比較結果是不同的，也就是欄位comparison為0的，所以推論如果有100筆資料的話，比較結果即為28/100，代表會有28筆比較結果不同，因此得出結論為：兩個資料的相似率大約為72%。但由於決策樹特徵數較少，只有5個，而且仔細觀察的話，當論文進度小於50時，且開會頻率為中的情況下，與論文進度大於等於50 時，且開會頻率為低的情況下，最終結果會是一樣的，因此推測如果特徵數一增加，那麼準確率就會下降。