DataMing HW2

姓名 : 吳建澄
學號 : NM6111035

Classification Problem : What kind of men do women prefer
Feature Information:
Age : 22-26
Handsome : 0-5 (5 is most handsome)
Height : 160-190
Body Fat% : 8-35
Crazy on work : 0, 1
Attitude : 0, 1
Wage : 25200-150000
Temper : 0-5 (5 is better temper)
Responsibility : 0, 1
Have a cat : 0, 1
Ma bao : 0, 1
Have a car : 0, 1
Have a house : 0, 1
Have a rich dad : 0,1
Jai Jai : 0,1
Label : 0(bad), 1(good)
Absolute right rule :
Score >= 7

Score rule:
Wage >=130000, Score + 5
Wage >=100000, Score + 3
Wage >=60000, Score + 1
Temper == 5, Score + 2
Temper == 4, Score + 1
Temper <= 2, Score -1
Responsibility == 1, Score + 3
Attitude == 1, Score + 3
Attitude == 0, Score -1
Ma bao == 1, Score -2
Ma bao == 0, Score + 1
說明 : 此分類問題為"哪種特質男生比較受女生喜歡"，在Feature Information中列出了此問題的所有feature資訊, 前半部為feature name, 後半部為feature值。而absolutely right rule也列在上面, 最後我做了一個加權的評分, 這些加分當中也有扣分的, 例如如果這個男生是"媽寶"則會-2分, 而評分超過7分及為"Good man"。

Generate Data:

我寫了一個簡單的程式來生成Data，Data生成總數為10000筆，為了讓good man跟bad guy的資料不要相差太多，原本想說大概5 分就可以稱為 "Good man" 了，但因為感覺現實中沒有那麼多好男人(1:1)，因此變更評分標準為 7 分以上才能稱為好男人(1:3)。

Result: 結果分成Decision Tree、naïve Bayes來討論，training set : test set = 8:2。

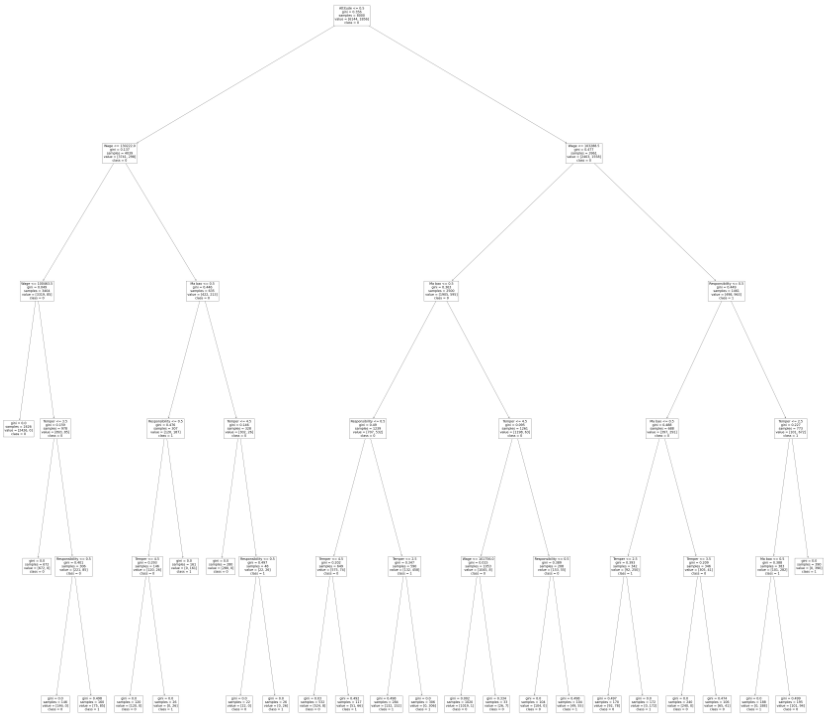
1: Decision Tree

```
accuracy: 0.999
Accuracy: 0.928
              precision    recall  f1-score   support

           0       0.95      0.95      0.95      1490
           1       0.86      0.86      0.86       510

    accuracy                           0.93      2000
   macro avg       0.90      0.91      0.91      2000
weighted avg       0.93      0.93      0.93      2000
```

上圖示跑完的結果，第一個accuracy是沒有限制model的節點所推出的結果，但是因為他節點數實在太深，導致輸出的圖片就算放大也看不到他的判斷依據是什麼，因此想說我只有5個Rules去評斷結果深度就放5試看看，得出的結果為0.928也是相當不錯。下圖是深度為5的Decision Tree，若看不清楚可參考附件(tree_d5.png)。



而從decision tree來看，由root往下分別是Attitude、Wage、Ma bao、Resbosibility、Temper，正好是absolutely right rule，代表說decision tree train出來的結果還算正確，

並沒有被其他較無相關的data搞混。但是我原本認為第一層應該是Wage > 13000, 畢竟這樣就5分了應該會切出一大半人, 但是它是由 是Attitude 來區分的。

2: naïve Bayes

Accuracy: 0.8175

	precision	recall	f1-score	support
0	0.81	0.98	0.89	1490
1	0.86	0.34	0.49	510
accuracy			0.82	2000
macro avg	0.84	0.66	0.69	2000
weighted avg	0.82	0.82	0.79	2000

上圖為跑完Naïve Bayes的結果, 可以發現結果跟Decision Tree相比好像沒那麼好, 可能是我設定absolutely right rule的方式是由加權機制所造成的, 例如說Wage > 130000有大機率是"Good man" 但是我還有其他feature 會造成逞罰扣分的機制, 因此就算有錢他的脾氣不好是個媽寶等等的因素, 可能會造成女生的印象大扣分。

總結:

整個作業的難處以及最特別的方式就是要如何產出Data了，雖然code都是參考上一屆學長的, 但從一開始訂定題目, 到設定哪些是absolutely right rule, 都是需要精心設計過的, 雖然此Data只是隨便設計好玩的(當然有些女生會看身高以及有無房車等), 一開始設定是只要那些有加分的feature全部交集起來就是好男人, 但這樣發現樣本太少, 以及在用Decision Tree 都是99%準確率, 因此才想出加分的機制去決定, 讓model添加一些不確定因素, 降低一下model的準確率, 進而觀察model和我預想中的是否一樣, 但是加了這個加分機制後, 若不做深度的限制, model的深度會深到我們無法去觀測(tree.png)。

而在建完data後, 用了decision tree、Naïve Bayes去train, 從結果來看decision tree的結果是比較的, Naïve Bayes則是較差的。