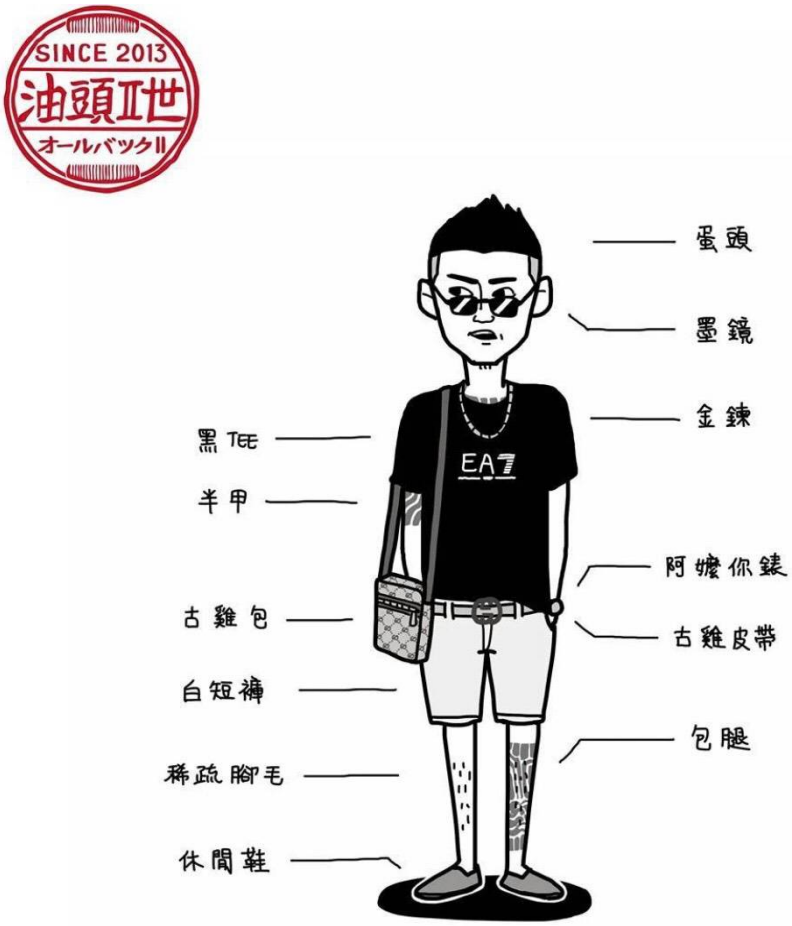
資料探勘報告Project 2

Classification



R26104047 張晏壬 (統研二)

老 師: 高宏宇 老師

目錄

1 目的 Abstract

本次作業主要利用多種不同之分類器（決策樹、邏輯斯迴歸、KNN、...），以分析在不同模型下對資料集的預測結果，而此次主要目標爲「利用民眾之個人基本資料，來預測此人是否爲具有前科之不良份子。」，資料設計的分析主要對象爲18歲以上之台灣民眾資料。

2 特徵設定 Setting

2.1 Input 特徵設計

- 性別（0 : 男生, 1 : 女生）
- 身高（設定男生平均身高：174 cm , 女生平均身高：161 cm，皆以高斯分佈生成）
- 體重（設定男生平均體重：64 kg , 女生平均體重：52 kg，皆以高斯分佈生成）
- 最高學歷（1：小學及以下, 2：國中, 3：高中職, 4：大專院校, 5：研究所及以上）
- 每星期平均抽幾根菸？（0 ~ 40根）
- 一星期平均喝幾瓶酒（0 ~ 10瓶，以均勻分配隨機生成酒瓶數）
- 刺青程度（0 ~ 10，0：沒有刺青, 10：刺滿全身）
- 吸食毒品的次數(0 ~ 15次)
- 髮色鮮豔程度（0 ~ 20，0：無染髮, 數字越高髮色越鮮豔）
- 平常愛穿的服飾品牌（1：NET, 2：H&M, 3：GU, 4：NIKE, 5：Addidas, 6：PUMA, 7：鬼洗, 8：CK, 9：GUCCI, 10：KENZO ）
- 去廟會的頻率（1：很少or幾乎不去, 2：偶爾去, 3：經常去, 4：每天報到）
- 身上最常攜帶的配飾（0：無, 1：耳環, 2：手錶, 3：佛珠, 4：金項鍊）
- 最喜歡的色彩（1：白, 2：紅, 3：橙, 4：黃, 5：綠, 6：藍, 7：紫, 8：黑）
- 平均每個月的收入（設定平均月收入爲44k，呈現高斯分佈）
- 台語流暢程度（0 ~ 10，數字越高，代表台語越流利）

2.2 Output 特徵設計

- 是否擁有前科（0：無, 1：有）

2.3 Absolute right rule

有前科紀錄者，需滿足A1 或A2 兩種情況之條件：

A1. 有前科紀錄(criminal_record = 1)：同時滿足1, 2, 3項特徵，且擁有4, 5, 6其中一項特徵

1. 一星期平均喝幾瓶酒>= 4
2. 刺青程度：>=2.5
3. 每星期平均抽幾根菸>=1
4. 去廟會的頻率：3：經常去, 4：每天報到
5. 平常愛穿的服飾品牌：7：鬼洗, 8：CK, 9：GUCCI, 10：KENZO
6. 髮色鮮豔程度>=3

A2. 有前科紀錄(criminal_record = 1)：擁有第6項特徵

6. 曾經吸食毒品的次數>=1

B. 無前科紀錄(criminal_record = 0)

- 不滿足前科紀錄條件

2.4 Setting 相關設定補充

1. 最高學歷（1：小學及以下, 2：國中, 3：高中職, 4：大專院校, 5：研究所及以上，設定爲多項式分佈，機率分別爲10%,12%,30%,40%,8%）
2. 設定每個人抽煙的機率爲42%，若有抽煙者，以指數分配隨機生成抽的菸數
3. 設定每個人吸毒的機率爲40%，若有吸毒者，以指數分配生成抽毒品的次數
4. 設定每個人刺青的機率爲42%，若有刺青者，以均勻分配隨機生成刺青程度
5. 性別、最喜歡色彩、平常愛穿的服飾品牌=> 呈現隨機分佈
6. 身高、體重受性別影響服從不同的高斯分佈
7. 最高學歷呈現多項式分佈，機率分別爲：小學及以下10%, 國中12%, 高中職30%, 大專院校40%, 研究所及以上8%
8. 去廟會的頻率呈現多項式分佈，機率分別爲：很少or幾乎不去50% , 偶爾去20%, 經常去15%, 每天報到15%
9. 髮色鮮豔程度與性別=> 正相關(女生染髮比例較高)

2.5 Noise干擾

1. 如果最高教育程度爲：小學及以下、國中、高中職，抽煙數量乘以2.5倍、刺青程度乘以1.5倍，吸毒之機率提高之60%
2. 如果性別爲男生，去廟會的頻率皆上升一級
3. 如果平均月收入>=60000者，抽煙數量乘以0.4倍

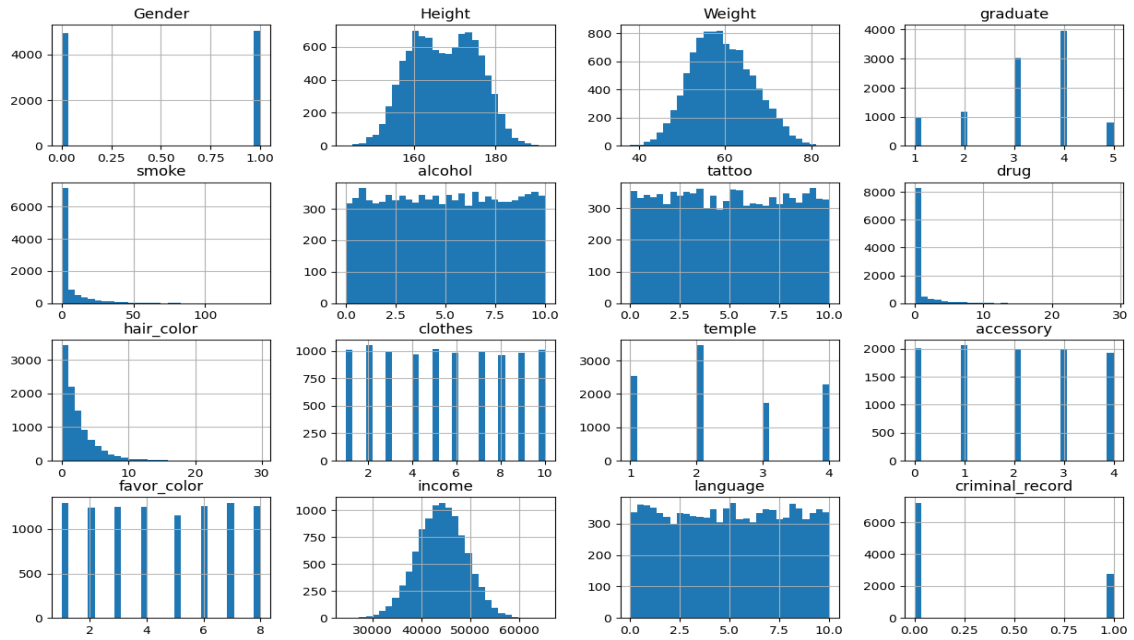
3 資料呈現Data appearance

經過第二節之設定完成後，我們加以根據設定之特徵生成資料，共生成10000筆共16項特徵之資料（最後一行「有無前科」即爲此次分類目標），如下所示：

性別	身高	體重	學歷	抽煙	喝酒	刺青	吸毒	髮色	衣服品牌	廟會	配飾	顏色	月收入	台語流暢度	有無前科
1.0	153.60	68.38	4.0	3.65	7.69	0.04	0.0	0.55	1.0	3.0	2.0	3.0	37847.0	6.94	0.0
0.0	183.45	68.12	2.0	0.00	6.58	2.61	0.0	0.93	3.0	4.0	4.0	3.0	41208.0	7.81	0.0
0.0	181.01	62.67	2.0	0.00	7.29	2.94	0.0	2.63	2.0	2.0	4.0	7.0	41520.0	4.85	0.0
1.0	165.22	46.36	1.0	0.00	0.53	5.66	0.0	2.04	3.0	2.0	0.0	4.0	51804.0	5.44	0.0
1.0	166.01	55.49	3.0	0.00	2.27	1.31	0.0	3.26	2.0	2.0	1.0	8.0	37784.0	5.26	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0.0	174.01	71.94	1.0	0.00	0.94	3.72	4.0	1.22	5.0	2.0	0.0	8.0	48675.0	5.65	1.0
1.0	161.88	55.42	4.0	17.84	6.38	0.91	0.0	0.61	5.0	1.0	3.0	4.0	42360.0	4.42	0.0
0.0	163.27	66.36	4.0	0.00	8.02	2.45	6.0	3.72	4.0	3.0	3.0	5.0	42663.0	8.73	1.0
0.0	172.82	62.21	5.0	5.94	8.77	7.94	0.0	0.03	7.0	4.0	4.0	6.0	45524.0	1.30	1.0
1.0	160.47	57.44	3.0	11.15	3.05	4.07	1.0	0.14	7.0	2.0	1.0	1.0	44769.0	6.83	1.0

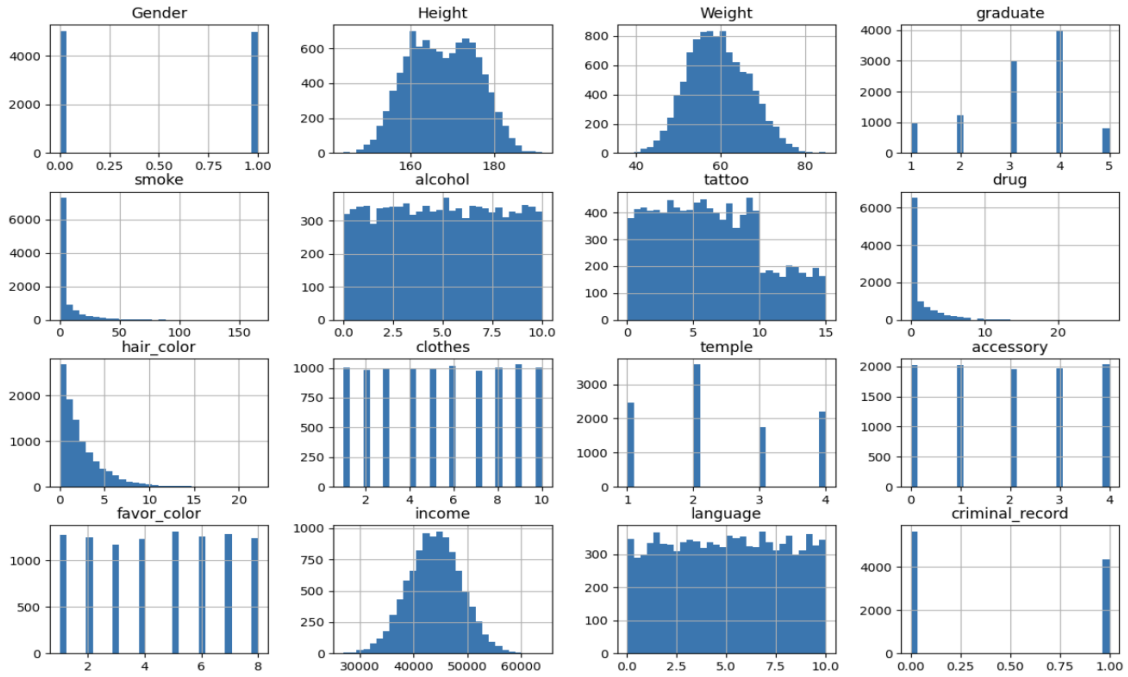
3.1 無noise干擾生成之資料

之後加以生成每個特徵之直方圖，如下圖所示，以檢視所有特徵之分佈情形，可觀察到生成的資料中，無前科與有前科的比例約爲3:1。



3.2 有noise干擾生成之資料

加入noise干擾後，行成之資料如下，其中有前科紀錄者大幅提升，有前科與無前科之比率來到11:9。



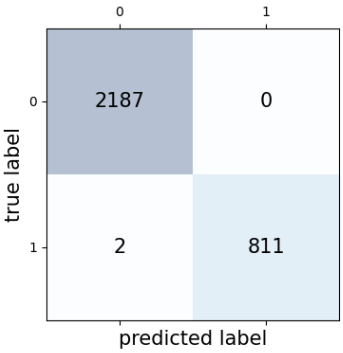
4　模型建立與評估Model and Evaluation

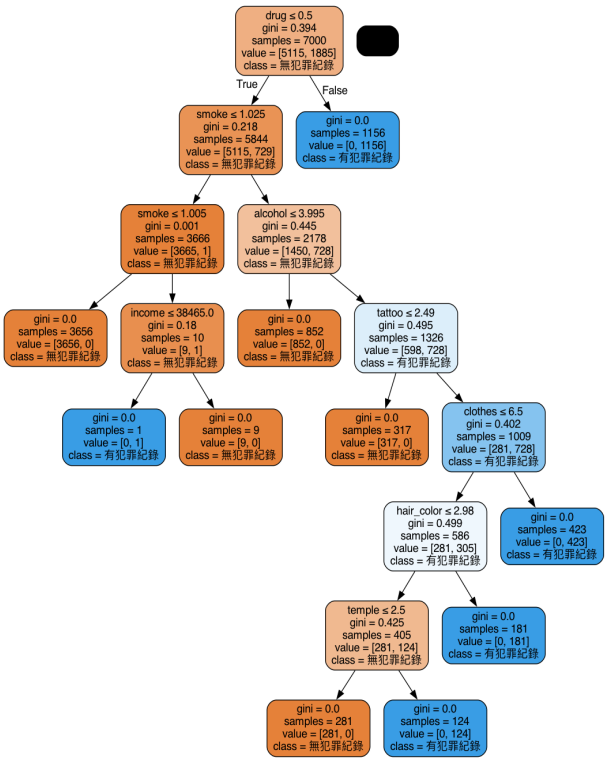
本次作業一共選取了下列幾種sklearn套件中的模型進行建模與比較，依序爲：決策樹、KNN、羅吉斯迴歸、高斯貝氏分類器，先將生成之資料使用sklearn套件中之utils.shuffle打亂排序，再使用train_test_split函數將資料分割成7:3的比例（7000筆訓練資料，3000筆測試資料），之後將資料帶入不同的模型，設定有無noise干擾，並在最終比較其模型分類之差別以找出其中關鍵因素。

4.1　決策樹Decision Tree

4.1.1　無noise干擾

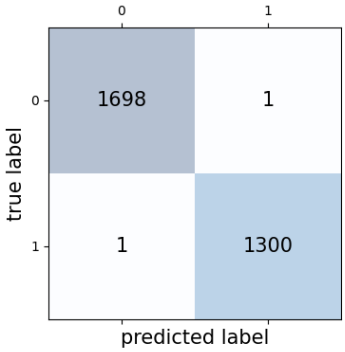
	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	2187
1.0	1.00	1.00	1.00	813
accuracy			1.00	3000
macro avg	1.00	1.00	1.00	3000
weighted avg	1.00	1.00	1.00	3000

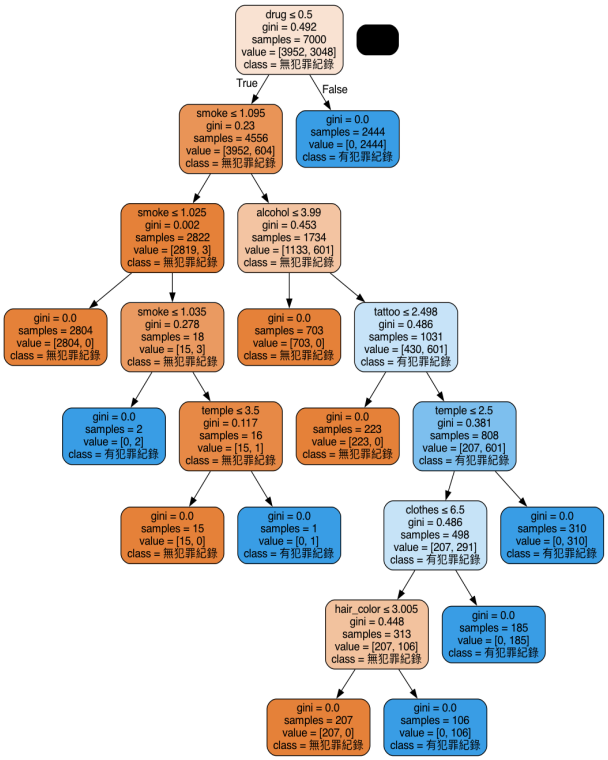




4.1.2 有noise干擾

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	1639
1.0	1.00	1.00	1.00	1301
accuracy			1.00	3000
macro avg	1.00	1.00	1.00	3000
weighted avg	1.00	1.00	1.00	3000

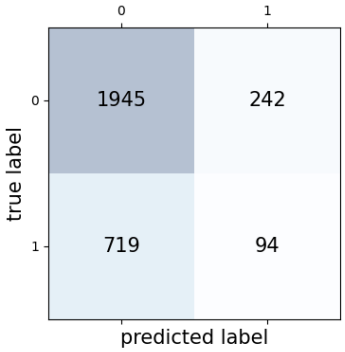




4.2 K-近鄰演算法 KNN

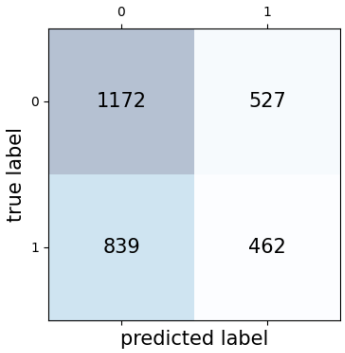
4.2.1 無noise干擾

	precision	recall	f1-score	support
0.0	0.73	0.89	0.80	2187
1.0	0.28	0.12	0.16	813
accuracy			0.68	3000
macro avg	0.50	0.50	0.48	3000
weighted avg	0.61	0.68	0.63	3000



4.2.2 有noise干擾

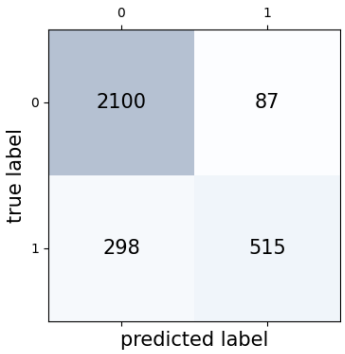
	precision	recall	f1-score	support
0.0	0.58	0.69	0.63	1699
1.0	0.57	0.36	0.40	1301
accuracy			0.54	3000
macro avg	0.52	0.52	0.52	3000
weighted avg	0.53	0.54	0.53	3000



4.3 羅吉斯迴歸 Logistic Regression

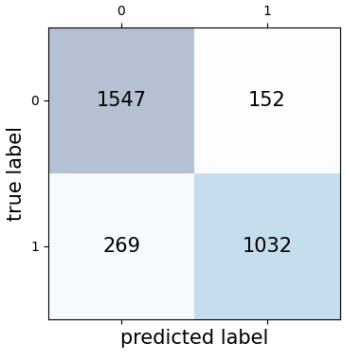
4.3.1 無noise干擾

	precision	recall	f1-score	support
0.0	0.88	0.96	0.92	2187
1.0	0.86	0.63	0.73	813
accuracy			0.87	3000
macro avg	0.87	0.80	0.82	3000
weighted avg	0.87	0.87	0.87	3000



4.3.2 有noise干擾

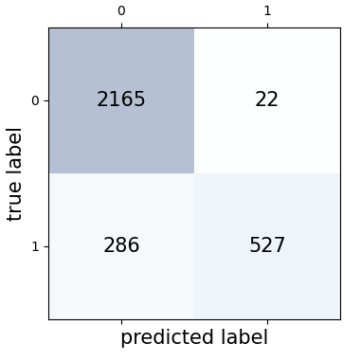
	precision	recall	f1-score	support
0.0	0.85	0.91	0.88	1699
1.0	0.87	0.79	0.83	1301
accuracy			0.86	3000
macro avg	0.86	0.85	0.86	3000
weighted avg	0.86	0.86	0.86	3000



4.4 高斯貝氏分類器 Gaussian Naive Bayes

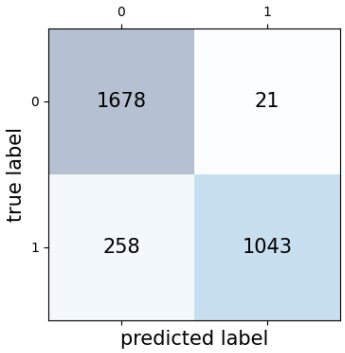
4.4.1 無noise干擾

	precision	recall	f1-score	support
0.0	0.88	0.99	0.93	2187
1.0	0.96	0.65	0.77	813
accuracy			0.90	3000
macro avg	0.92	0.82	0.85	3000
weighted avg	0.90	0.90	0.89	3000



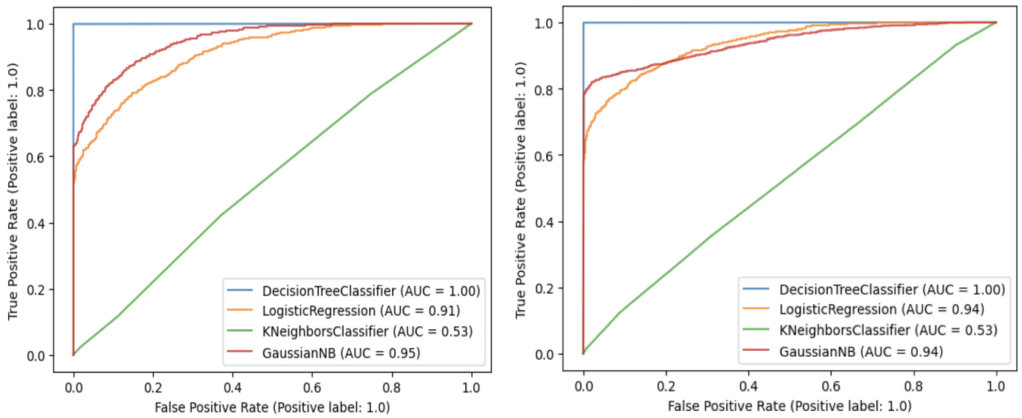
4.4.2 有noise干擾

	precision	recall	f1-score	support
0.0	0.87	0.99	0.92	1699
1.0	0.98	0.80	0.88	1301
accuracy			0.91	3000
macro avg	0.92	0.89	0.90	3000
weighted avg	0.92	0.91	0.91	3000



4.5 有無noise干擾總比較—ROC曲線

畫出有無noise干擾之資料帶入模型後之ROC曲線，並進行比較。



5 結論Conclusion

此次報告花費最多的時間實爲「資料之設計」，對於每一項特徵變數經過多次調整後（嘗試減少類別特徵，增加連續型特徵），決策樹的準確率依舊是高得嚇人，永遠均逼近100%，唯一有試過可行的辦法爲將目標變數「有無前科」的條件再加上：當滿足2.3節之所有absolute right rule情況下，確實有前科的機率不是100%，而是70%，將可有效降低決策樹的準確率，但想想感覺違反absolute right rule的原則，因而作罷。

而綜觀有無加入噪音干擾，與各種不同的模型輸出結果可以發現，不管有無noise干擾下，「決策樹」在此資料中的分類皆爲最佳表現，原因爲資料進行設計時，特徵與特徵之間並無特別強之關聯性，決策樹擅長處理不相關的特徵，儘管後續使用noise強化關聯性，但猜測設定的還是不夠強；而所有模型中「KNN」則皆爲最差，KNN模型表現最差的原因可能爲因該資料中有眾多類別型資料，而KNN主要運算方法爲計算每點特徵之「距離」，其對資料型別要求較高，較適合連續型特徵。

從決策樹畫出的枝葉圖可發現，模型對於較單純之absolute right rule(吸毒次數大於等於1者⇒前科=1)確實能在一開始進行準確之分類，而對於設定之「需同時滿足之特徵」，也能觀察到其在枝葉圖之節點相較於「擁有其他特徵」落於較上層，但值得注意的是在無noise干擾情況下之枝葉圖，其中一項節點的分類標準爲「月收入是否大於38465元」，此爲在absolute right rule無設定之條件，但加入noise之後，強化了「月收入(income)」與「抽菸數量(smoke)」兩個特徵之間的關聯，在枝葉圖中「月收入」爲節點分類標準的情形就消失了。

除了最佳與最差之模型外，此次也加入了最常被用於二分類的「羅吉斯迴歸模型」與「高斯貝氏分類器」，高斯貝氏模型假設所有的特徵都是相互獨立且服從常態分配，透過總比較之AUC曲線可發現，在無noise情況下，高斯貝氏分類器效果會略勝於羅吉斯迴歸，但在加入noise之後，可發現因特徵關聯性被強化，而導致高斯貝氏分類效果降低，變成與羅吉斯迴歸效果相同。

此次報告讓我了解到，每個資料適合的模型不同，若要分析的準確，則必須考量到資料中的特徵是否滿足模型之假設，才能理解爲何模型在資料中準確率低落的原因，進而思考針對特徵進行轉換或修改模型之參數。