資料探勘報告 Project 2

Classification

自知之明男性評價分類器

P76121657 鄭翊宏

老師:高宏宇 老師

中華民國 2022 年 11 月 12 日

摘要

本作業受網路上諸多評價男性價值啟發，目的為建立男性對自己的自知之明，本作業位男性標準分為三類，分別是帥哥、普男及其他，本初期先以主觀判斷為標準去決定規則，最後在將規則套用在測試者身上，本作業會著重於評估各分類器對於主觀判斷的評價分類是否準確，並進行分析，最後最後提出本專題發現的問題及未來目標

目錄

1. 特徵設計 Data Setting

1.1 input 特徵設計

-身高(設定男生平均身高 174cm ，以高斯分布生成)

-體重(設定男生平均體重 64 kg ，以高斯分布生成)

-最高學歷(1:國中,2:高中職,3:大專院校,4:研究所,5:博士)

-年收入(1:學生,2:0~50 萬,3:51 萬~100 萬,4:101 萬~150 萬,5:151 萬~200 萬,6:201 萬以上)

-抽菸喝酒(1:不菸不酒,2:喝酒不抽菸,3:抽菸不喝酒,4:抽菸喝酒)

-顏值(1~10,請參考 Figure 1.1)

-日常穿衣風格(1:舒適為主,2:方便為主,3:好看為主)

-交往經驗(1:0 位,2:1~3 位,3:4~6 位,4:7 位以上)

-運動習慣(1:不運動,2:一周 1~2 次,3:一周 3~4 次,4:一周 5 次以上)

-刺青程度(0~10,0:沒有刺青,10:全身刺青)

-擁有房屋數(1:0 棟,2:1 棟,3:2 棟,4:3 棟以上)

-約砲經驗(1:不約,2:偶爾約,3:常常約,4:每天約)

1分男
油腻结块的头发
父母一方可能不是人类
五官大小失衡
皮肤粗糙
气质猥琐
上辈子得罪了全宇宙？
身材短胖
因为只能与右手作伴，这类男人已经濒危
2分男
女人和他啪啪时会恶心呕吐，要吐很多次才能习惯
五官有严重问题
会像1分男一样被认为是非人类
人靠衣装在他身上失效
很可能是一名杀马特
3分男
虽不吓人，但没人想多看一眼
物质条件越差恋爱经历越逼近0
无大问题小问题不断的脸
¥
0
经历
最常见的吊丝男，最大梦想是成功逆袭和抱得女神.
4分男
无硬伤也无亮点的
大众脸
路人甲
过目即忘的长相.
一提起"接盘侠""老实人"，最可能是他们.
在生活中随可见的4分男，女神一招手可能会来一堆.
5分男
五官周正
无硬伤
一眼看去说不上喜欢 但一定不会被讨厌.
也不至于太帅被人觊觎
或许会因为既不丑也不帅，而缺乏存在感吧
5分注定是个不上不下的分，但和前面几种吊丝比，5分男还是很动人的.
6分男
比5分男多一些帅气之处
恭喜开始进入帅哥行列！
可能在身高、五官或气质上有缺陷而不能成为7分
6分男可能有不少的女吊丝暗恋
7分男
在大学一定是校草
是女生们暗恋明恋的对象~
新潮发型
双眼皮
五官精致
耳环也不违和
高鼻子
小脸
可能爱化妆
身高180+，身形健美
娱乐圈小鲜肉、摄影师帅哥、gay时尚达人
经常出自7分男
8分男
自信
只要出现，
会引起骚动
和偷拍
靠脸吃饭不成问题.
少女们把8分男的照片设成背景，
8分男已不仅拥有颜值和身材还拥有
比7分男更高的辨识度
可能是某个家喻户晓的男明星
9分男
9分男可能上辈子救了全宇宙才换来这外貌
无法用言语形容他的帅
因为可能只是幻想
他可能是银幕偶像，可能是国民情人
神话级别的存在
可能已故多年但依然被铭记
可能是跨多个领域的天才
?
10分男
颜值
身材
名望
智商
社会地位
努力
这些他都有
知乎 @心记录

1.2 output 特徵設計

-男生評價(1:高價值 2:中等, 3:其他)

1.3 Absolute right rule

評價為高價值者須擁有 1,2,3 其中一項,或同時滿足 4,5,6,7 所有特徵:

1. 擁有房屋數>=3(2 棟以上)
2. 顏值>=7
3. 年收入>=5(151 萬 up)
4. 身高>=180cm
5. 體重<=75KG
6. 運動習慣>=3(一周 3 次 up)
7. 抽菸喝酒<=2(不可抽菸)

評價為中等者為不滿足高價值條件者中，須擁有 1,2,3 其中一項,或同時滿足 4,5,6,7,8 所有特徵

1. 擁有房屋數>=2(1 棟以上)
2. 顏值>=6
3. 年收入>=4(51 萬 up)
4. 身高>=170
5. 體重<=70
6. 運動習慣>=2(一周 2 次 up)
7. 抽菸喝酒<=2(不可抽菸)
8. 約砲經驗=1(不約)

評價為其它者為皆不滿足高等與中等價值者之餘下男性

1.4 相關設定細項

-身高(設定男生平均身高 174cm ，以高斯分布生成)

-體重(設定男生平均體重 64 kg ，以高斯分布生成)

-最高學歷(1:國中,2:高中職,3:大專院校,4:研究所,5:博士，設定為多項式分布，機率分別為 5%,30%,45%,15%,5%)

-年收入(1:學生,2:0~50 萬,3:51 萬~100 萬,4:101 萬~150 萬,5:151 萬~200 萬,6:201 萬以上，設定為多項式分布，機率分別為 20%,10%,50%,15%,5%)

-抽菸喝酒(1:不菸不酒,2:喝酒不抽菸,3:抽菸不喝酒,4:抽菸喝酒，設定每人喝酒的機率為 70%,抽菸的機率為 30%,故機率分別為 21%,49%,9%,21%)

-顏值(1~10，平均分數為 5，以高斯分布生成)

-日常穿衣風格(1:舒適為主,2:方便為主,3:好看為主，呈現隨機分布)

-交往經驗(1:0 位,2:1~3 位,3:4~6 位,4:7 位以上，設定為多項式分布，機率分別為 30%,50%,15%,5%)

-運動習慣(1:不運動,2:一周 1~2 次,3:一周 3~4 次,4:一周 5 次 up，設定為多項式分布，機率分別為 30%,40%,20%,5%)

-刺青程度(0~10,0:沒有刺青,10:全身刺青，設定每個人刺青的機率為 35%，有刺青者隨機生成刺青程度)

-擁有房屋數(1:0 棟,2:1 棟,3:2 棟,4:3 棟以上，設定為多項式分布，機率分別為 40%,40%,15%,5%)

-約砲經驗(1:不約,2:偶爾約,3:常常約,4:每天約，設定為多項式分布，機率分別為 70%,25%,3%,2%)

1.5 Noise 干擾

1.如果最高學歷<=2(高中職以下)，則刺青程度乘以 1.5 倍，抽菸喝酒機率改為(抽菸 50%,喝酒 85%，故機率變為 7.5%,42.5,7.5%,42.5)，且如果最高學歷>=4(研究所以上)，則年收入往上升 1 個等級，擁有房屋數有 50%的機率上升一個等級

2.如果日常穿衣風格為 3(好看為主)，則顏值上升一個等級

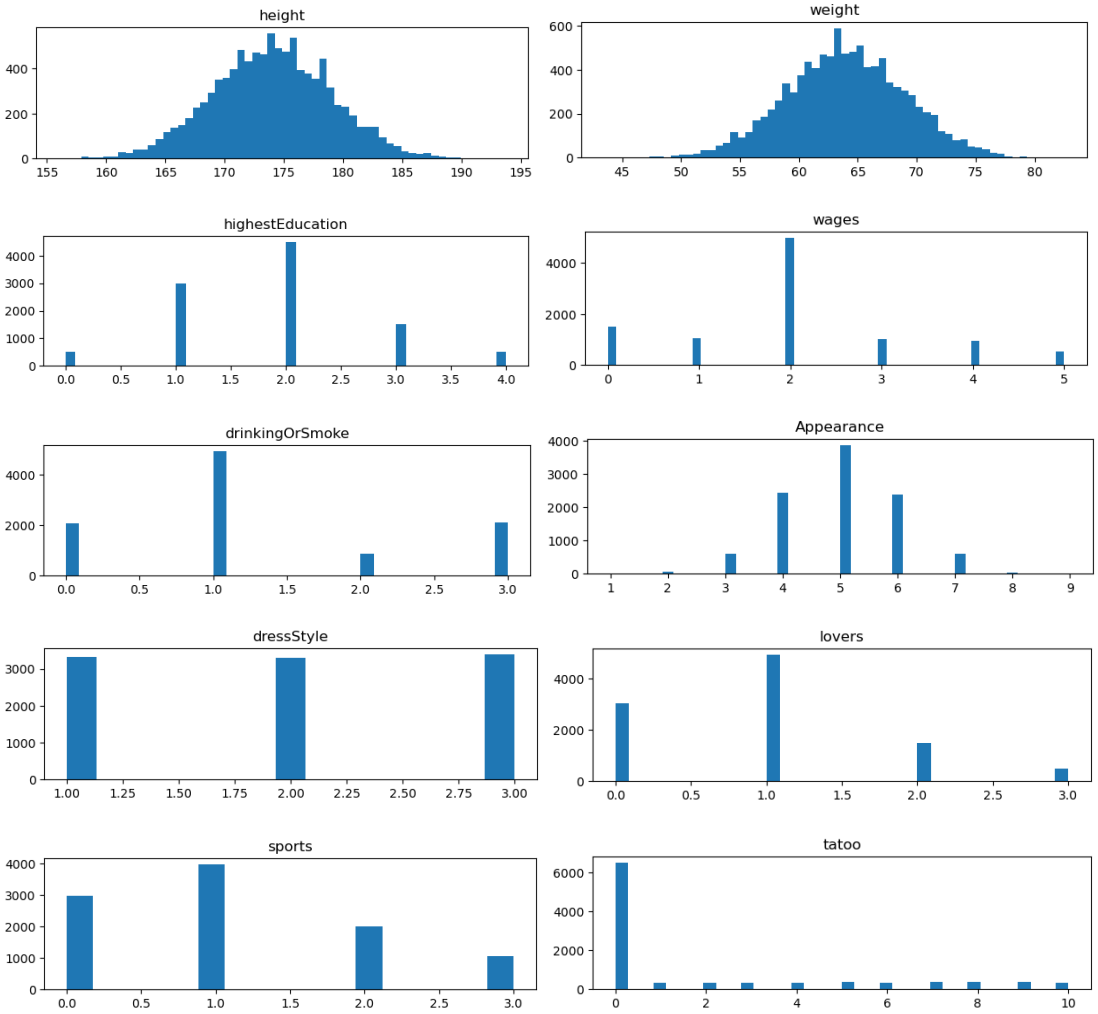
3.如果顏值>=7，則約砲經驗有 30%的機率上升一個等級

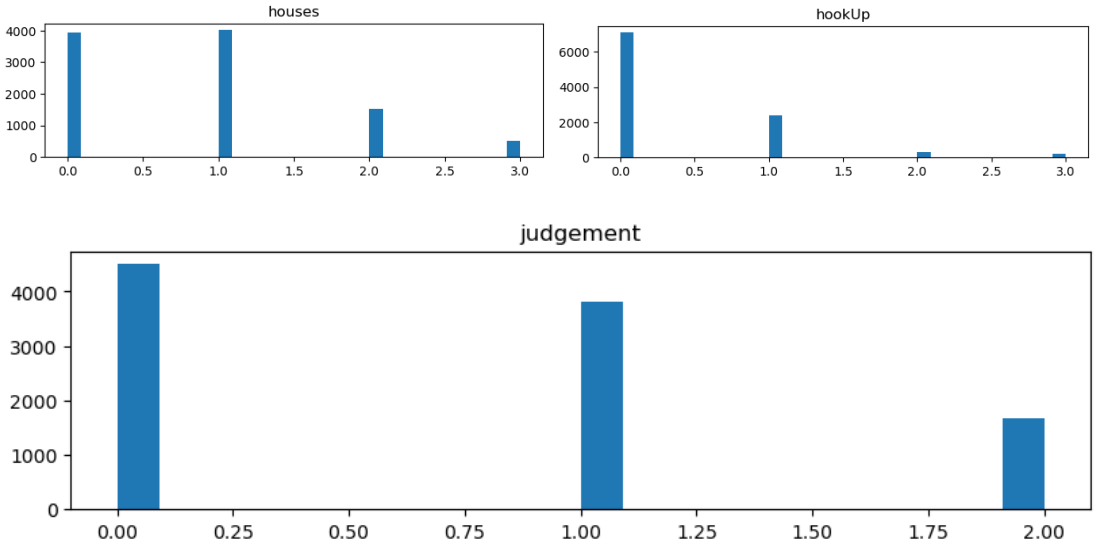
2.資料呈現 Data appearance

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 13 項特徵資料(最後一項為標籤項極為分類目標)，如下所示:

2.1 無 noise 干擾生成之資料

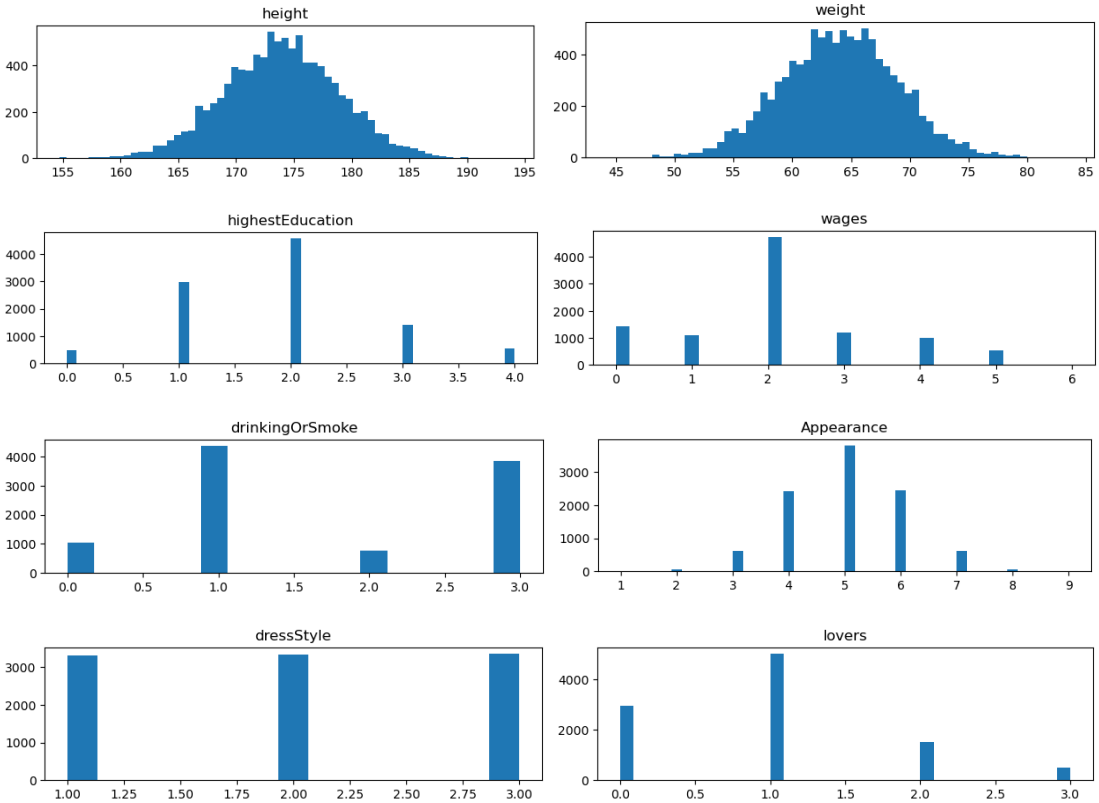
之後加以生成每個特徵之直方圖，如下圖所示，以檢視所有特徵之分布情形，可觀察到生成的資料中，低中高比例分別為 45% ,38%,17%

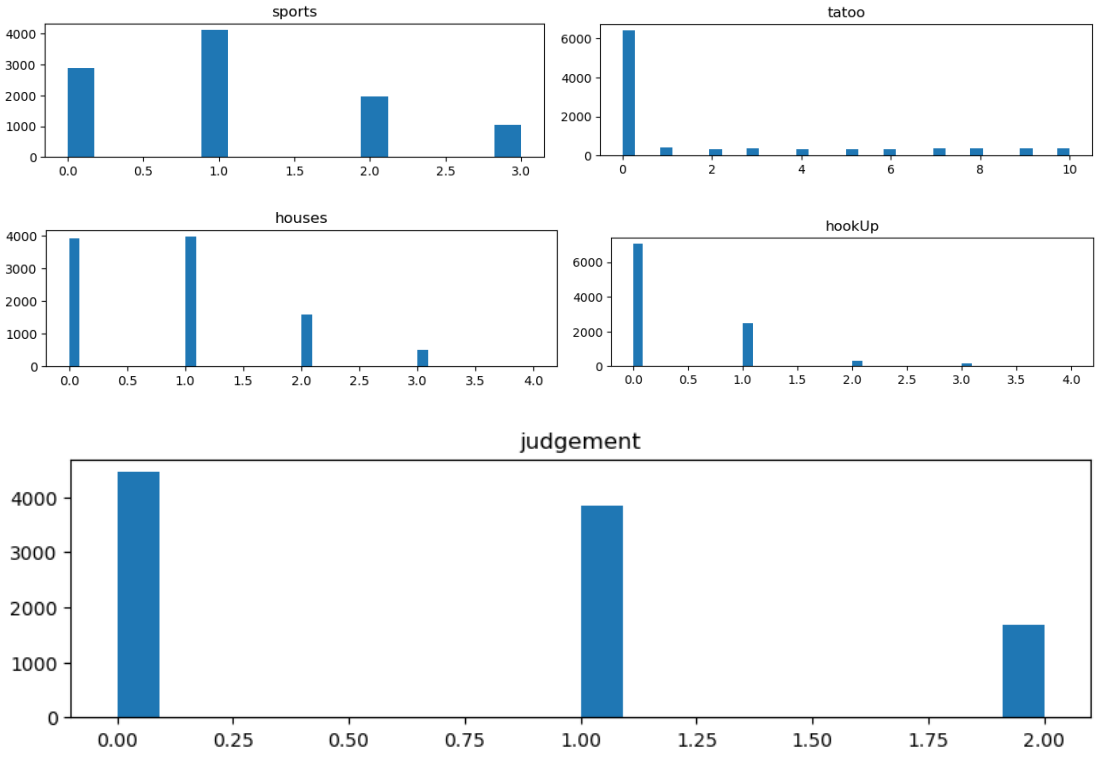




2.2 有 noise 干擾生成之資料

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布情形，可觀察到生成的資料中，低中高比例分別為 45% ,38%,17%




sports
tatoo
houses
hookUp

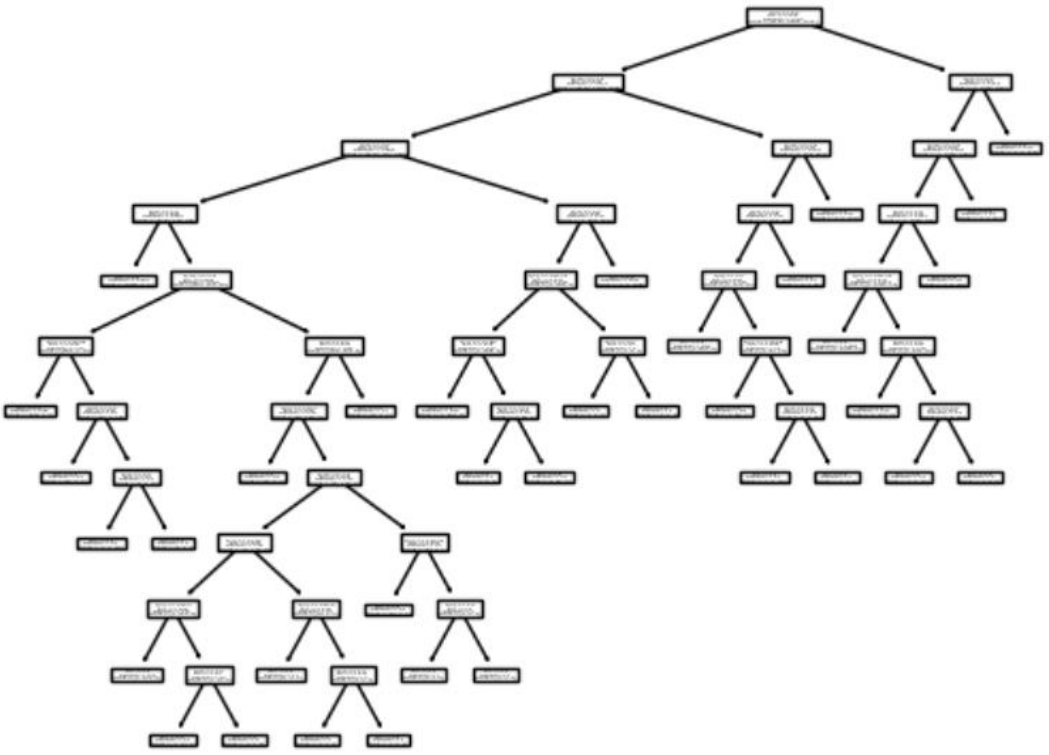
judgement

3.模型建立與測試 Model building/testing

本次作業一共選取了下列幾種 sklearn 套建中的麼行進行建立與評估，分別為，先將生成之資料亂序排列之後再將資料切成 7:3 (7000 訓練資料與 3000 筆測試資料)，之後再將資料帶入不同的模型，設定有無 noise 干擾，並在最終比較其分類之差別以找出其中關鍵因素

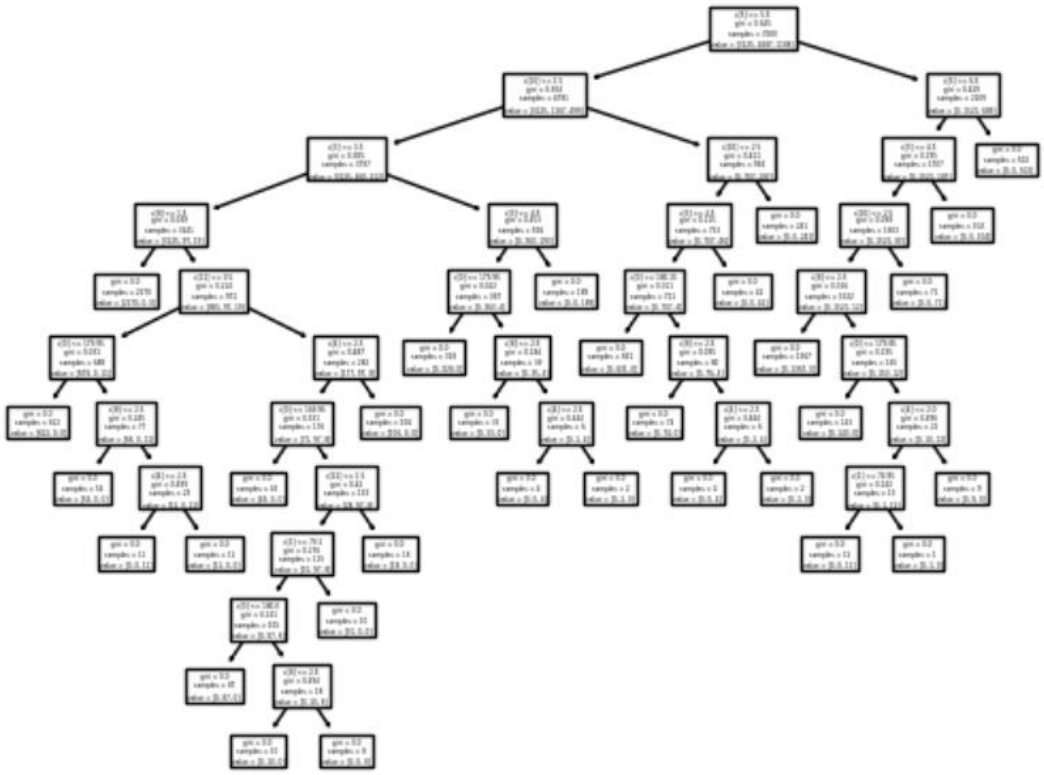
3.1 Decision tree

3.1.1 無 noise 干擾



準確率為 0.998

3.1.2 有 noise 干擾



準確率為 0.9997

3.2 SVM

我們比較 3 種不同的 kernel 函數對於資料的準確率

3.2.1 rbf(Radial basis function)

```
NoNoise: 0.8633333333333333
Noise: 0.8696666666666667
```

3.2.2 linear

```
NoNoise: 0.699
Noise: 0.696
```

3.2.3 sigmoid

```
NoNoise: 0.45866666666666667
Noise: 0.438
```

3.3 KNN

我們比較 3 種不同的 neighbo 數對於資料的準確率

k=5(default)

```
NoNoise: 0.6266666666666667
Noise: 0.631
```

k=3

```
NoNoise: 0.6366666666666667
Noise: 0.6303333333333333
```

k=7

```
NoNoise: 0.6256666666666667
Noise: 0.6166666666666667
```

3.4 Gaussian Naive Bayes

```
NoNoise: 0.8626666666666667
Noise: 0.859
```

4.結論 Conclusion

4.1 討論與總結

此次報告花費最多的部分為特徵設計，從思考一般人衡量之標準到設計盡量符合現實之比例，我發現不管是否加入 noise，得出來的結果比例相差無幾，這反映出男性價值相對不容易出現流動，而主要體現價值的身高體重、顏值部分基本上屬於先天便決定好的，而錢財部分一部分取決於家世背景也是屬於先天的，唯一能改善的只有提升自己的工作能力，但在現環境薪資待遇普遍低落的情況下能有好薪水的少之又少，造就了男性低價值的比例接近一半，而不是平均的狀況。

在分類標準上本次報告共使用了 4 種不同的分類器分別為 Decision tree, SVM, KNN, Gaussian Naive Bayes，在準確率上 Decision tree 明顯高於其他分類器，推測因為 Absolute right rule 的規則屬於 if else 取向的簡易劃分規則，並且各個特徵之間的關聯性別沒有很大，因此 Decision tree 可以很好的劃分出這些區間，因此準確率最高，另外在這些分類器中平均表現最不好的是 KNN，估計是因為特徵之間的關連不大，並且判斷條件會出現 or 的判斷導致 KNN 取鄰近值的方法並沒有辦法很好的判斷出結果，故而平均準確率最低

4.2 未來展望

在未來我們期望可以利用更加準確的統計而不是預設的機率來建立個人模型，並且可以利用機器學習的方法來判斷顏值。