機器學習期末報告

公投分析

1.動機:  
我們的動機起源於台灣這次1124的九合一大選，除了地方公職人員的結果外，公投的結果也是一大熱門話題，而且剛好我們又要做機器學習的期末報告，所以我們就想說可以試著用機器學習的方法來分析這次的公投結果，我們想看看究竟什麼因素會影響公投結果，或是甚麼因素其實影響力沒有這麼大。

2.方法:  
我們以一個村里為一筆資料的單位(因為以一個鄉鎮的話，全台灣只有300多個，資料筆數太少，而全台灣大概有7.8000個村里，資料筆數比較夠)。我們去找出這些鄉鎮的各種人口資料，例如人口數或是年齡等等的，然後以這些人口資料去當機器學習的特徵，然後以公投結果當作目標，然後我們分別用三種不同的模型去跑機器學習。在這次這麼多公投案中，我們想要分析第十三案:是否同意東奧以台灣去參賽，以及第十案:是否同意婚姻是限於一男一女。

3.資料與特徵:

我們所有人口資料都是出自於內政部，然後所有的公投資料都是出自於中選會。我們總共選用的特徵有7種:戶人口數、居住情形、教育程度、婚姻情形、人口總數、性別比、年齡情形。

定義7特徵:

戶人口數:我們從內政部的人口資料得到一戶分別是幾人的各種戶數，我們統計成三群1-4人一戶、5-9人一戶、10人以上一戶，而我們把就把三群的戶數對人口數做加權平均，第一群的權重是3(我們覺得三人的小家庭應該會佔多數一點)，第二群的權重是7(單純取中位數)，第三群的權重是10(10人的家庭應該是大多數)，這樣做後得到的加權平均就是我們所謂的戶人口數，可以約略視為平均的戶人口。

居住情形:我們從內政部的人口資料得到獨居的戶數與群居的戶數(只要兩人以上一起住就算)，然後我們群居的戶數除以獨居的戶數，得到群居獨居比，就是所謂的居住情形。

教育程度:我們從內政部可以得到各種教育程度的人數，我們把所有教育程度類型按照同等教育程度歸類為:不識字、小學畢業、國中畢業、高中畢業、大學畢業、碩士畢業、博士畢業，然後分別給1-7分的權重，再去做加權平均，最後會得到一個數值，可以說是代表這個村里的教育成讀，越大就表示平均教育程度越高，就是所謂的教育程度。

婚姻情形:我們從內政部得到已婚人口與未婚人口，我們把已婚人口除以未婚人口，得到已婚未婚比，就是所謂的婚姻情形。

人口總數:我們從內政部得到一個村里的人口總數。

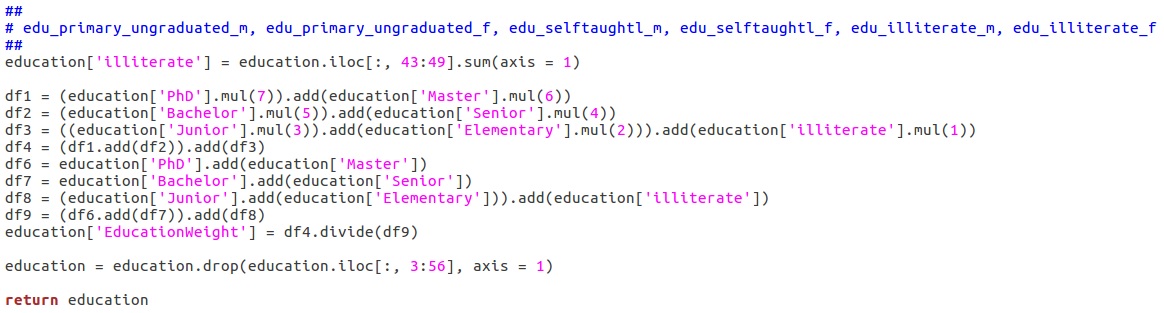
性別比:我們從內政部得到男生的人口數與女生的人口數，然後把男生的人口數除以女生的人口數，得到男女比，就是所謂的性別比。

年齡情形:我們從內政部得到各年齡的人口數，我們統計成四群，分別是18-38歲、39-59歲、60-80歲、81歲以上，然後作加權平均，第一群到第四群的權重分別就是1-4，最後會得到一個加權平均，可以約略視為一個村里的平均年齡，這就是所謂的年齡情形。

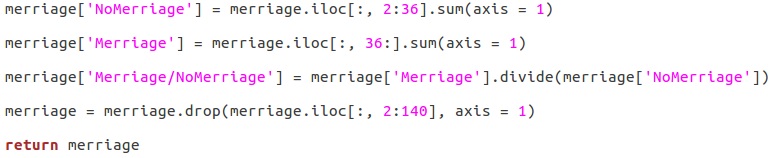
我們的目標就是公投結果的同意率。

4.資料處理:

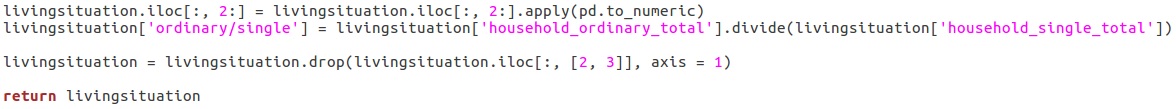
以下資料都是擷取了code的一部分



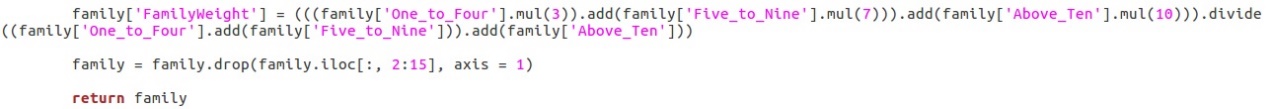
上圖是處理教育程度的程式，先統整出各教育程度的人數再做加權平均



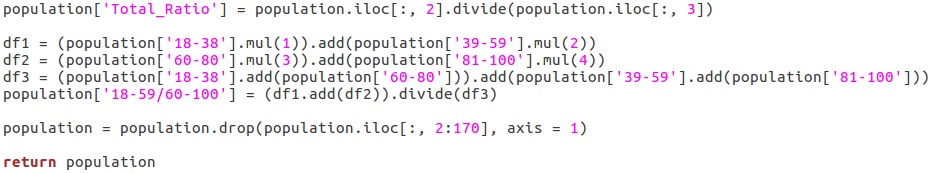
上圖是處理婚姻狀況的程式，先統整出未婚和已婚或曾已婚的人數再做人數比



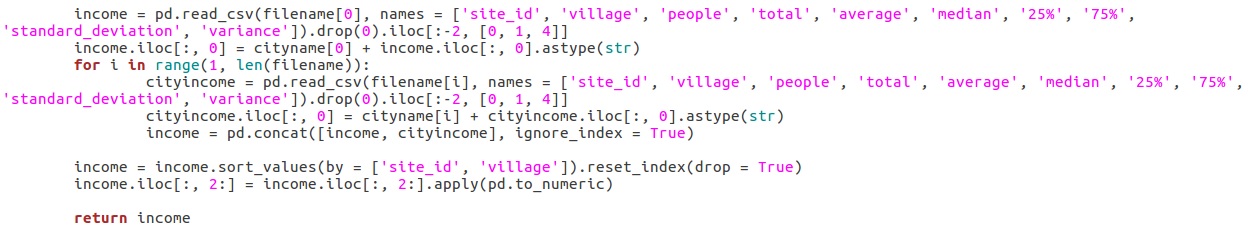
上圖是處理居住情形的程式，算出群居與獨居的比數



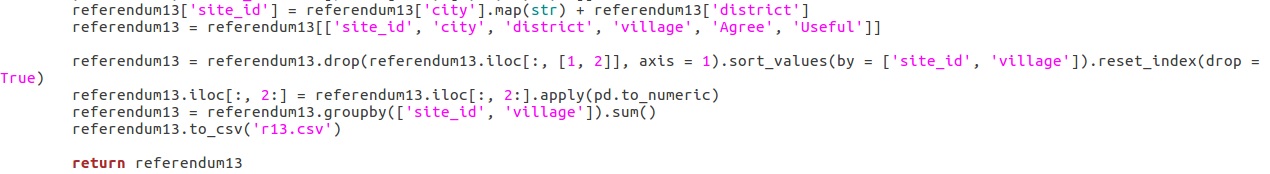
上圖是處理戶人口數的程式，會先將戶人口數分區間後再做加權平均



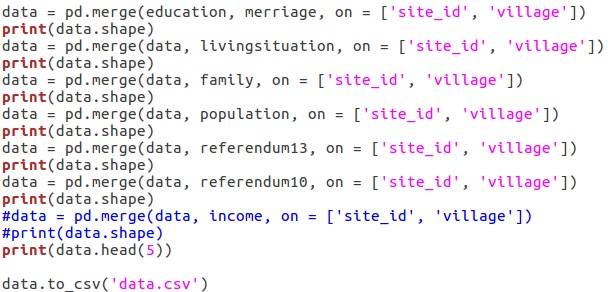
上圖是處理人口數的程式，會先將年齡分層再做加權平均，其中column名為”18-59/60-100”的就是加權平均，並非指18-57歲和60-100歲的人口比



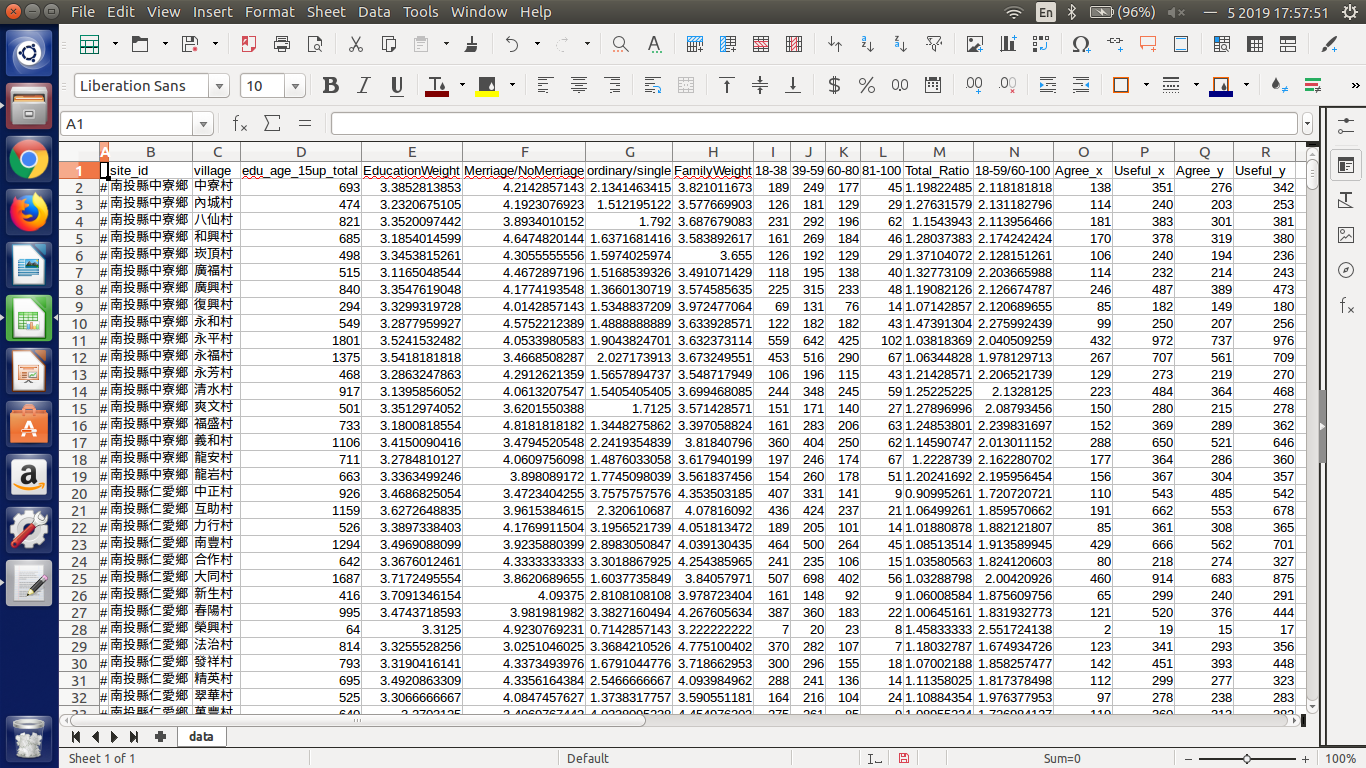
上圖是處理收入的程式，會統整出各鄉鎮市的平均收入，但經過各個model測試過後相關性太低所以最後捨棄。



上圖是處理公投結果的程式，13和10案的程式相同，都是抓出各鄉鎮市有效同意票和總有效票



上圖是將各個讀進來且處理好的資料統整成一個data，而且只留下每個資料都有的鄉鎮市，其中收入是後來不採用的



上圖便是統整完後輸出的data，其中"18-59/60-100"是指各年齡區間的加權平均，並非18-59歲和60-100歲的人口比數

5.三種模型:

我們使用的三種的模型分別為Decision tree、Knn、Regression，以下是這三種模型的簡介與特點。

Decision tree:

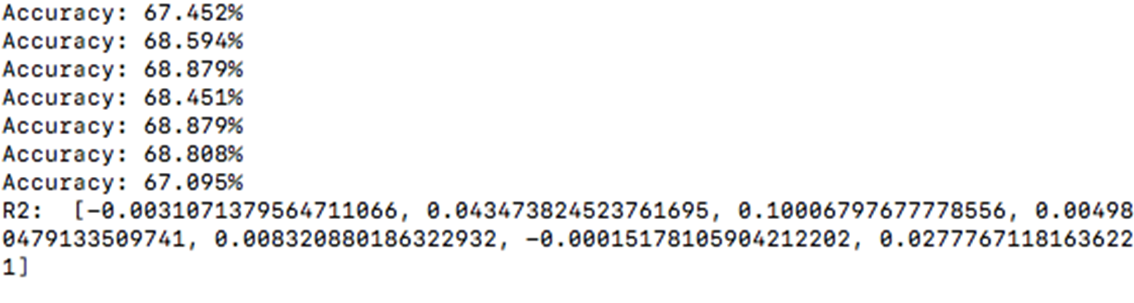
Regression:

Knn:

6.執行結果:

分析第十三案結果:

Regression: r2 of each features



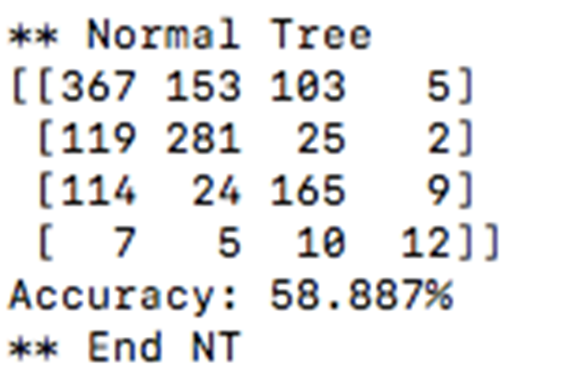
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Total**  **number** | **Education**  **weights** | **Marriage**  **situation** | **Household**  **situation** | **Family**  **weights** | **Sex**  **ratio** | **Age**  **weights** |
| -0.00310 | 0.04347 | 0.10006 | 0.00498 | 0.0083 | -0.00015 | 0.02777 |

Regression: r2 of total features



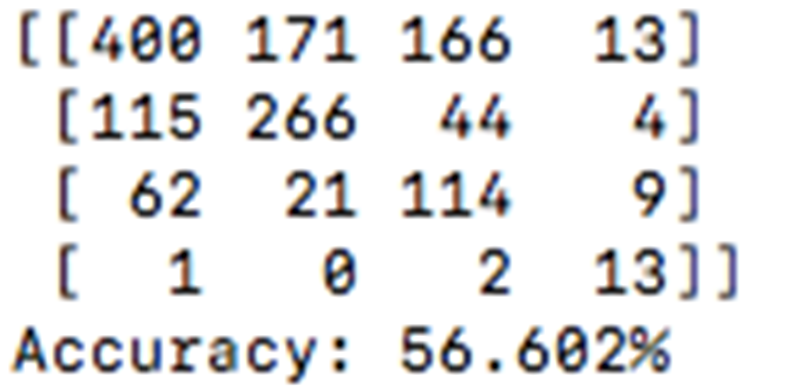
R2:0.24663

Decision tree: Accuracy of total features



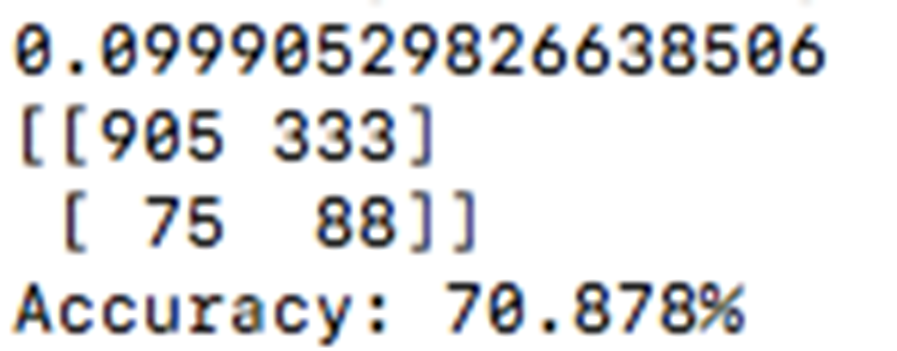
Accuracy: 58.887%

Knn: Accuracy of total features



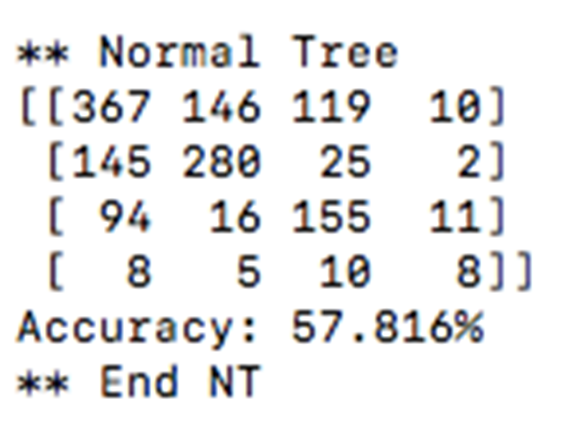
Accuracy: 56.602%

Regression: r2 of three features(marriage, family, age)



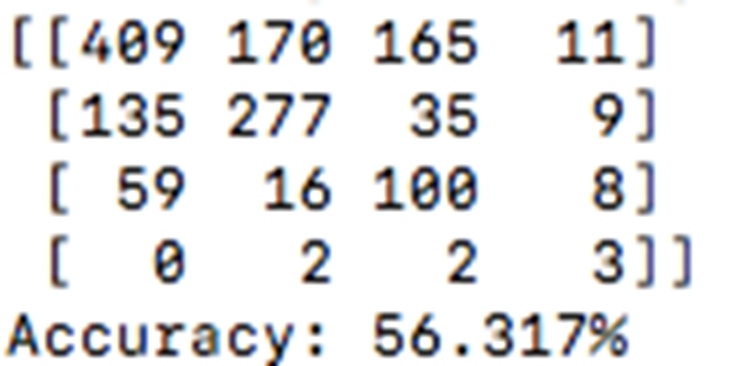
R2:0.0999

Decision tree: Accuracy of three features(marriage, family, age)



Accuracy:57.816%

Knn: Accuracy of three features(marriage, family, age)



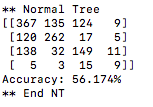
Accuracy:56.317%

Regression: r2 of three features(marriage, family, education)



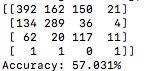
R2:0.2189

Decision tree: Accuracy of three features(marriage, family, education)



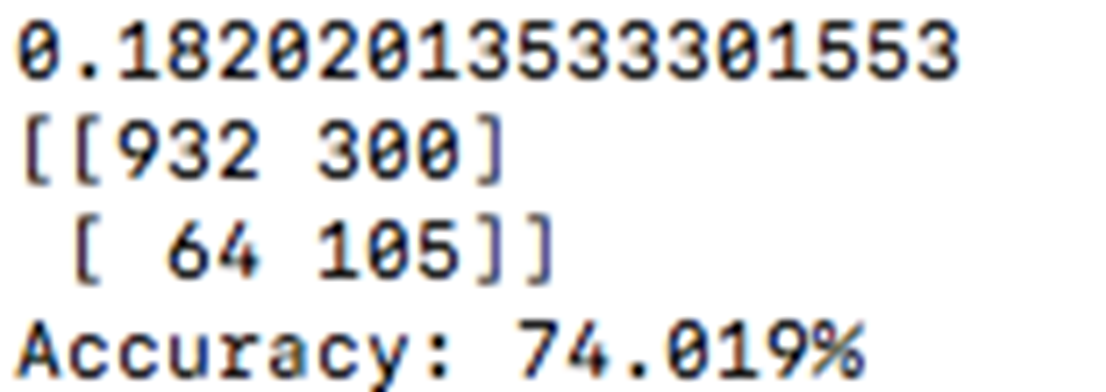
Accuracy:56.174%

Knn: Accuracy of three features(marriage, family, education)



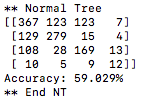
Accuracy:57.031%

Regression: r2 of three features(marriage, age, education)



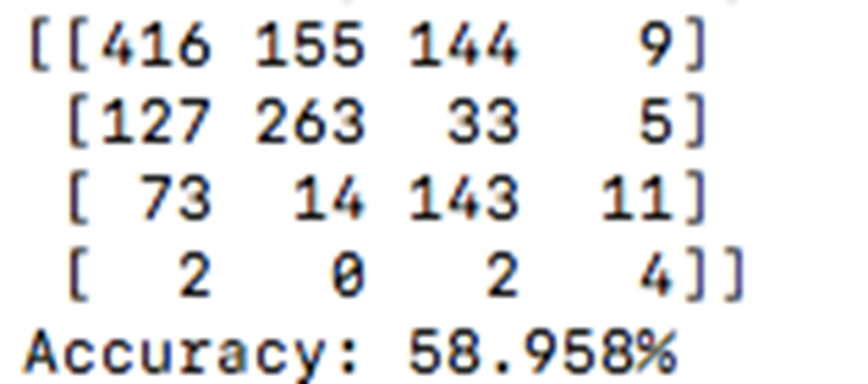
R2:0.1820

Decision tree: Accuracy of three features(marriage, age, education)



Accuracy:59.029%

Knn: Accuracy of three features(marriage, age, education)



Accuracy:58.958%

小結論:

每個特徵: 準確度不是太好

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Total**  **number** | **Education**  **weights** | **Marriage**  **situation** | **Household**  **situation** | **Family**  **weights** | **Sex**  **ratio** | **Age**  **weights** |
| -0.00310 | 0.04347 | 0.10006 | 0.00498 | 0.0083 | -0.00015 | 0.02777 |

所有的特徵: 還不錯,但不是真的到很好

|  |  |  |
| --- | --- | --- |
| **Regression: r2** | **Decision tree: Accuracy** | **Knn: Accuracy** |
| 0.24663 | 58.887% | 56.602% |

三種特徵組合:

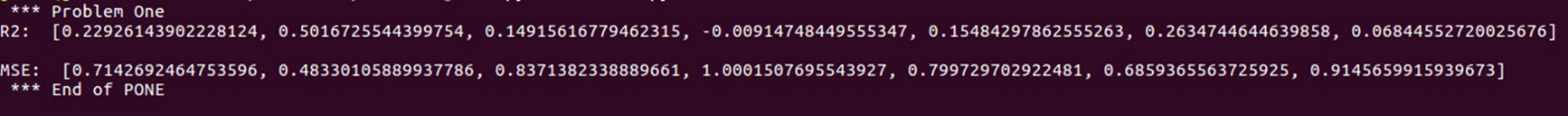
最好的組合:marriage, family, education

|  |  |  |  |
| --- | --- | --- | --- |
|  | **marriage, family, age** | **marriage, family, education** | **marriage, age, education** |
| Regression | 0.0999 | 0.2189 | 0.1820 |
| Decision tree | 57.816% | 56.174% | 59.029% |
| Knn | 56.317% | 57.031% | 58.958% |

分析第十案結果:

我們這案只使用的兩個模型，我們並沒有使用knn模型，因為

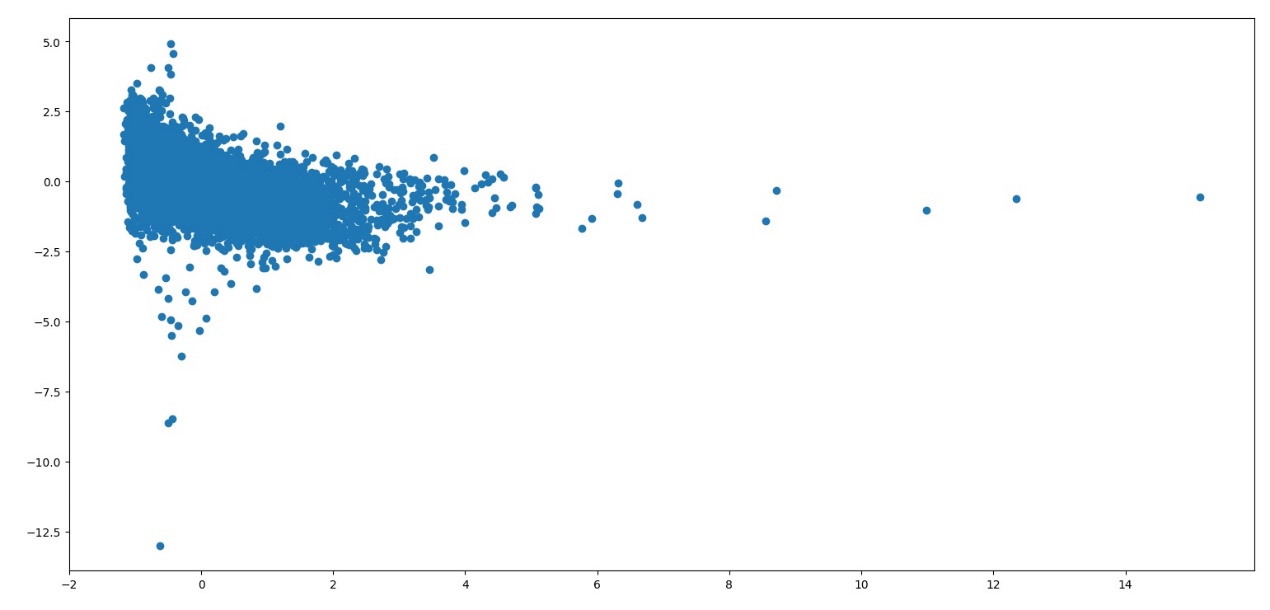
Regression: r2 of each features

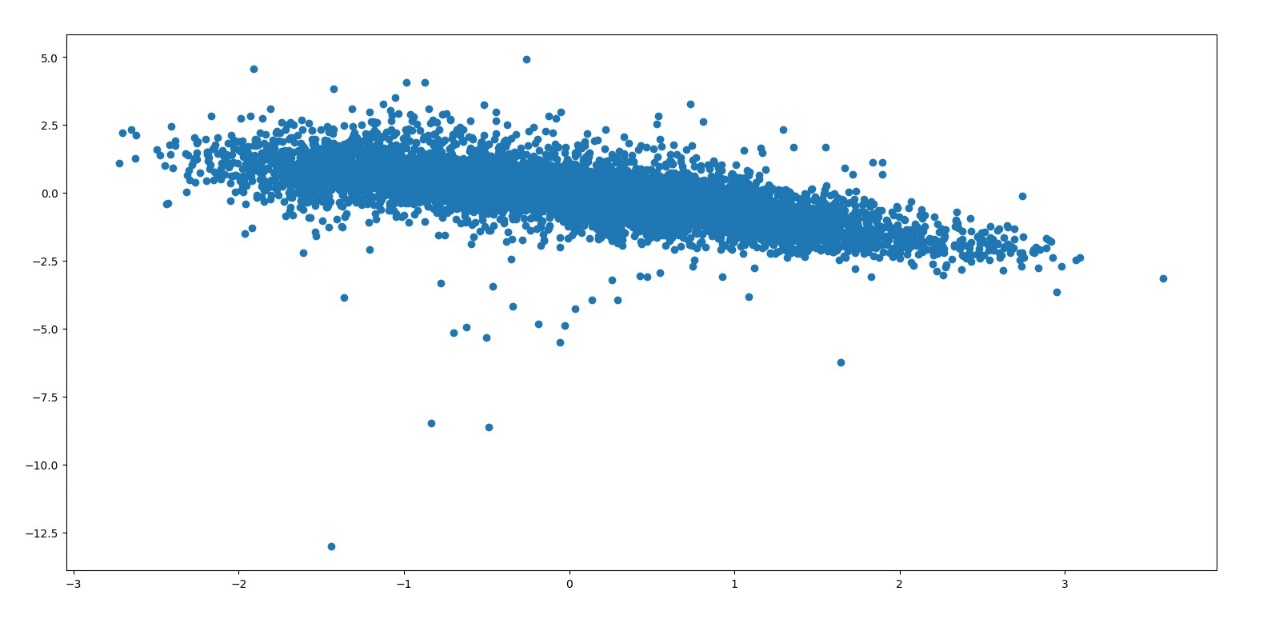


|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Total**  **number** | **Education**  **weights** | **Marriage**  **situation** | **Household**  **situation** | **Family**  **weights** | **Sex**  **ratio** | **Age**  **weights** |
| 0.22926 | 0.50167 | 0.14915 | -0.00914 | 0.15484 | 0.26347 | 0.06844 |

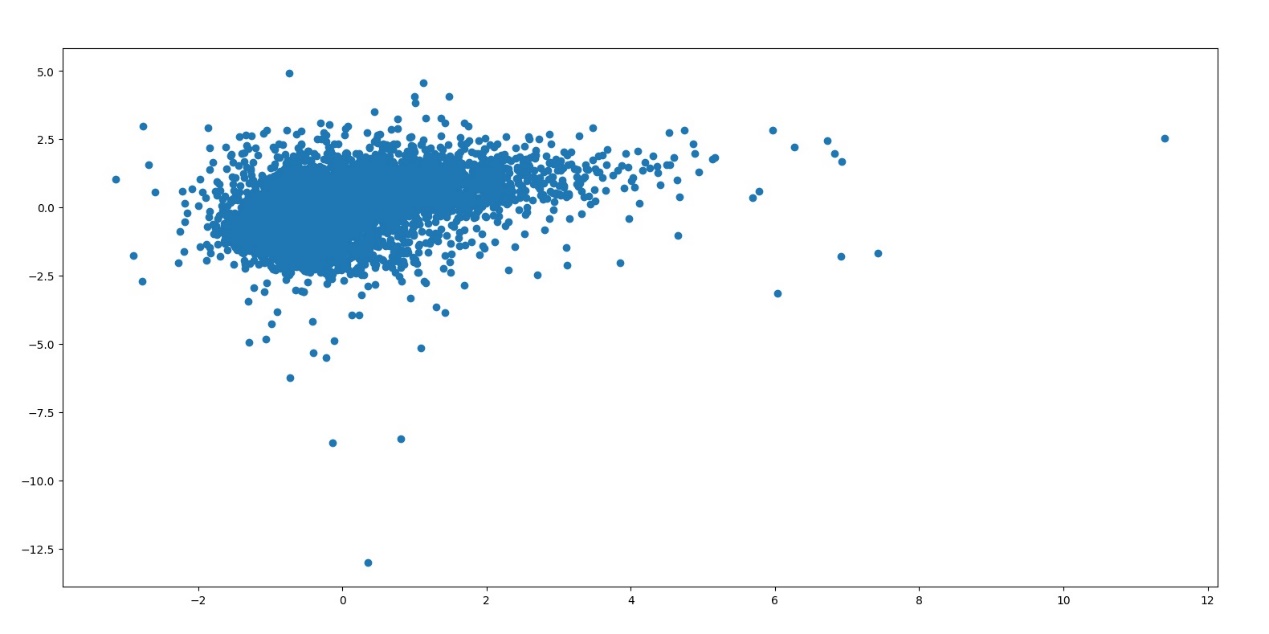
趨勢圖(各種特徵對公投同意率作圖):

總人口數

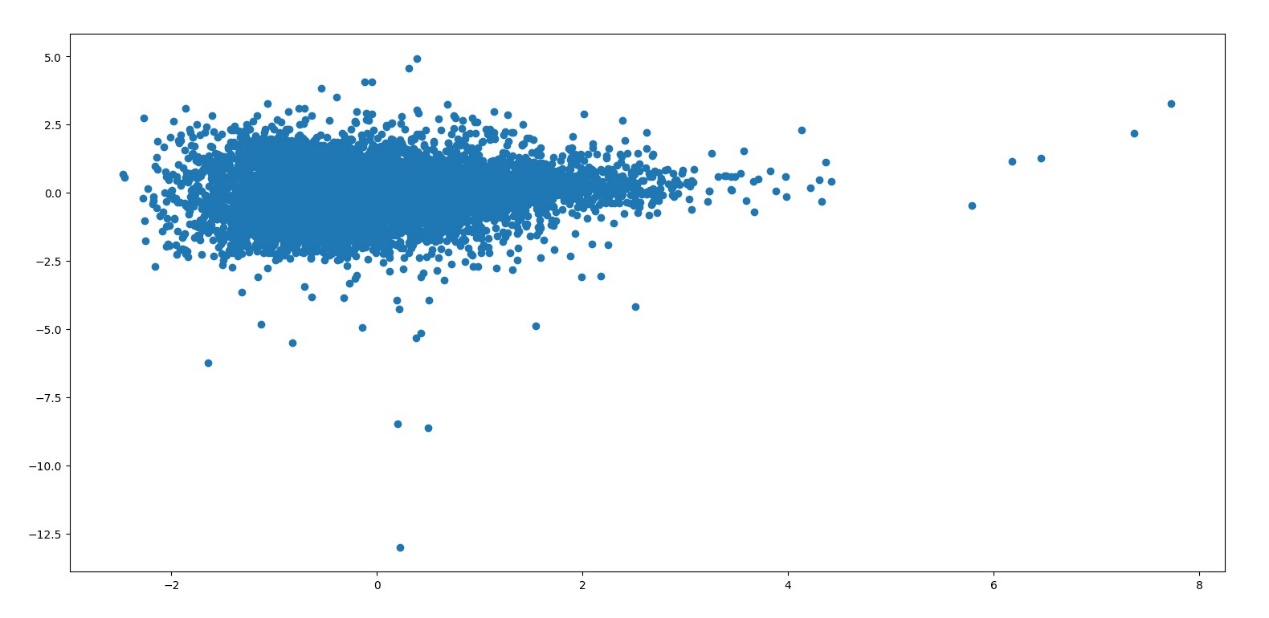


教育程度

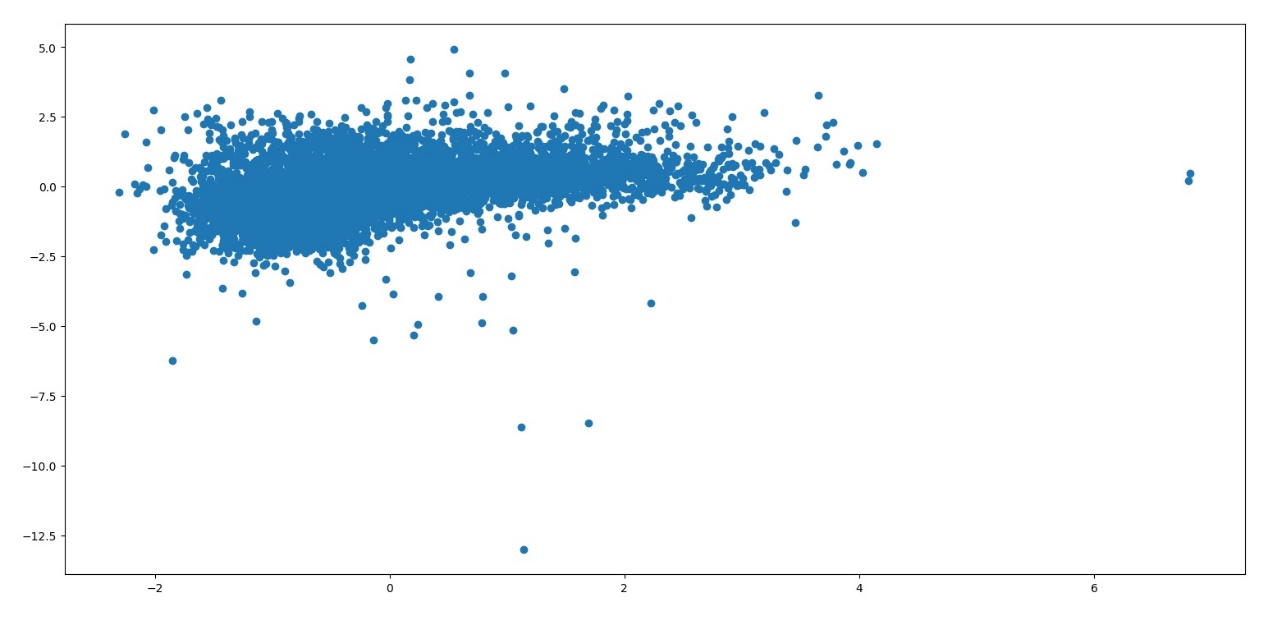
婚姻情形



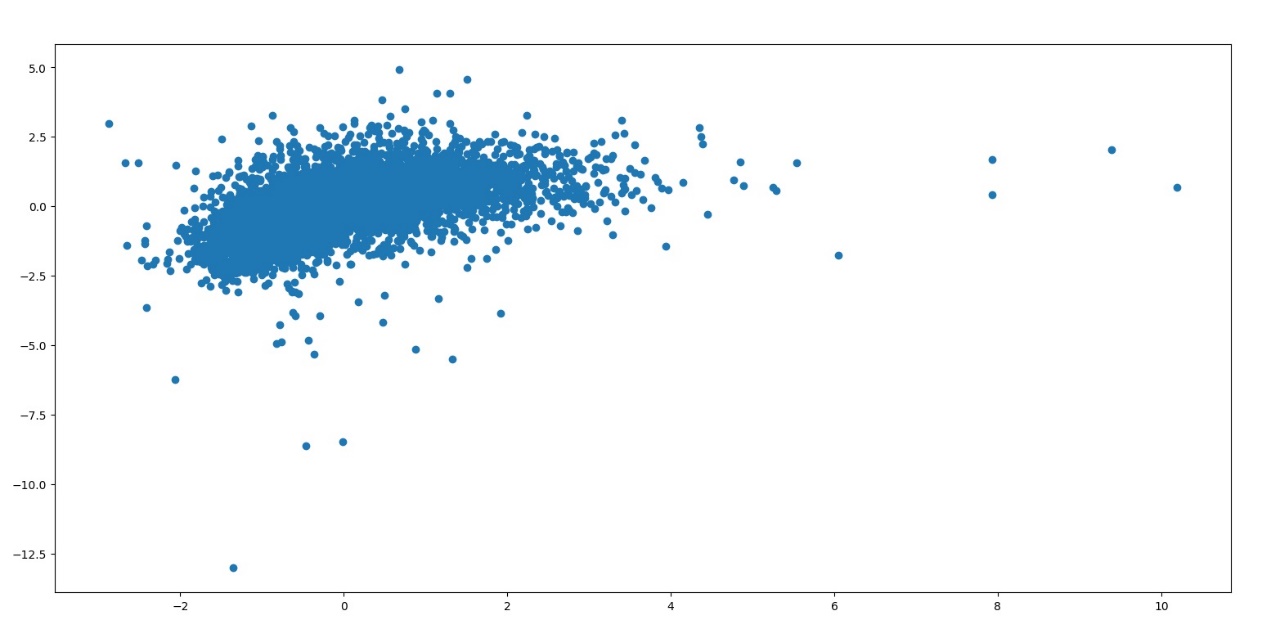
戶人口數



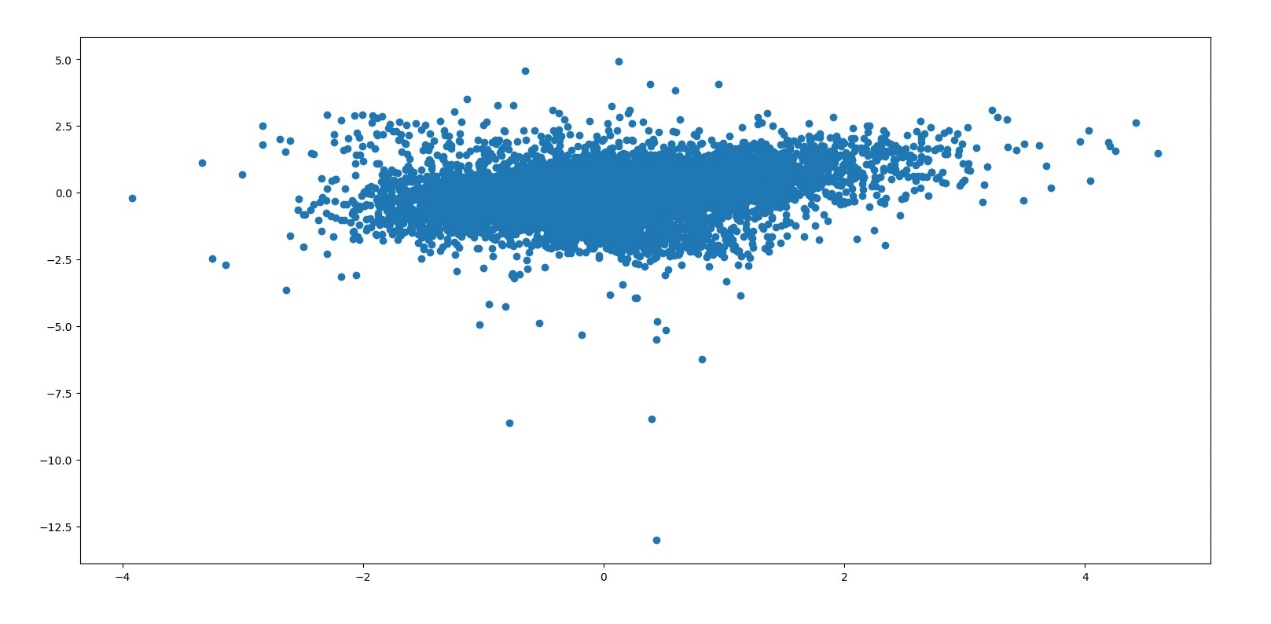
居住情形



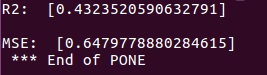
性別比



年齡情形

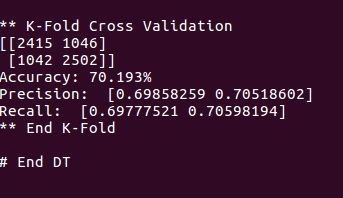


Regression: r2 of total features



R2:0.43235

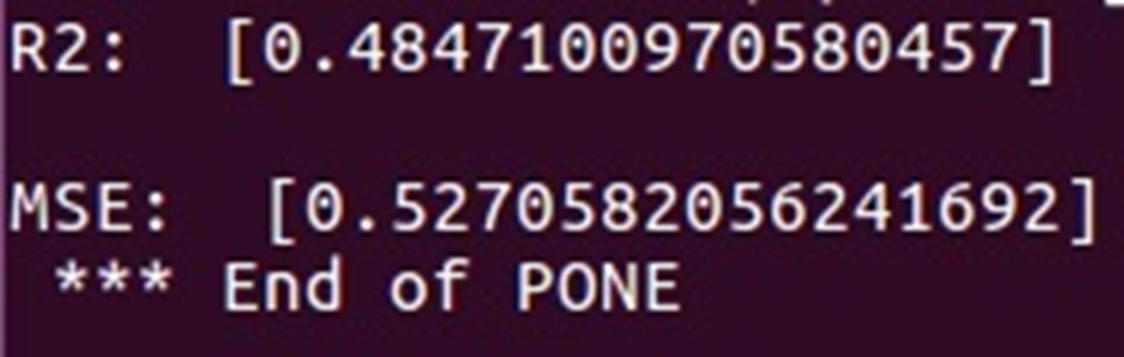
Decision tree: Accuracy of total features



Accuracy: 70.193%

Regression: R2 of three features(total, education, sex ratio)

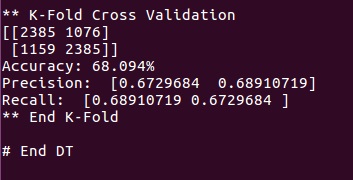
(R2 top three)



R2:0.484

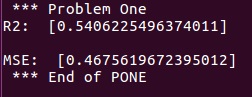
Decision tree: Accuracy of three features(total, education, sex ratio)

(R2 top three)



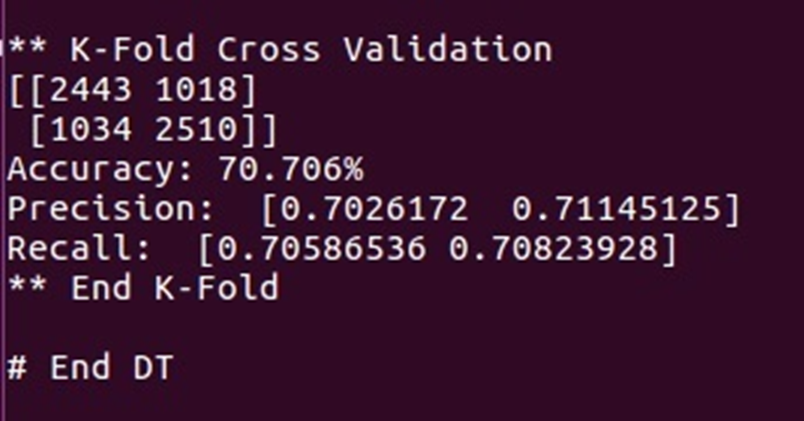
Accuracy:68.094%

Regression: r2 of three features(marriage, age, education)



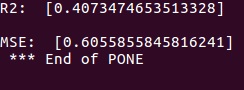
R2:0.4675

Decision tree: Accuracy of three features(marriage, age, education)



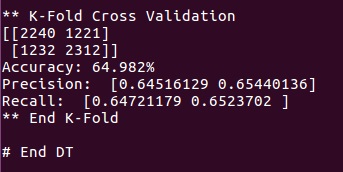
Accuracy:70.706%

Regression: r2 of three features(marriage, age, total)



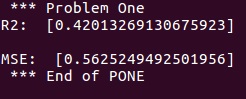
R2:0.40734

Decision tree: Accuracy of three features(marriage, age, total)



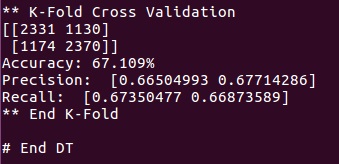
Accuracy:64.982%

Regression: r2 of three features(marriage, age, sex ratio)



R2:0.42013

Decision tree: Accuracy of three features(marriage, age, sex ratio)



Accuracy:67.109%

小結論:

所有特徵: 不錯

|  |  |
| --- | --- |
| **Regression: r2** | **Decision tree: Accuracy** |
| 0.43235 | 70.193% |

三個特徵群: 最好的: marriage, family, education

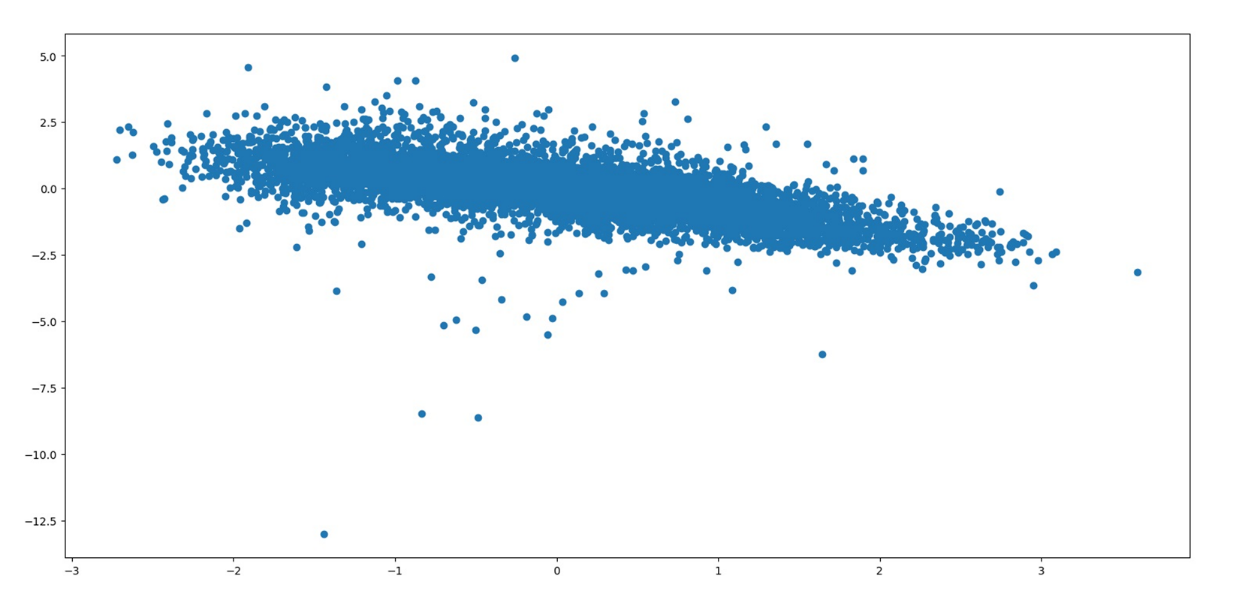
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **total, education, sex ratio** | **marriage, age, education** | **marriage, age, total** | **marriage, age, sex ratio** |
| Regression | 0.484 | 0.4675 | 0.40734 | 0.42013 |
| Decision tree | 68.094% | 70.706% | 64.982% | 67.109% |

對每一個特徵來講，相關程度最高的是Education weights

前三名分別為: Education weights, Sex ratio, Total number

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Total**  **number** | **Education**  **weights** | **Marriage**  **situation** | **Household**  **situation** | **Family**  **weights** | **Sex**  **ratio** | **Age**  **weights** |
| 0.22926 | 0.50167 | 0.14915 | -0.00914 | 0.15484 | 0.26347 | 0.06844 |

教育程度對投票同意率作圖



從上表可以看出，就數據面來說教育程度的r2其實高達0.50167，而且就趨勢圖來看，也可以滿明顯的看出是有一定的趨勢，就這張圖來說，我們可以說教育程度越高是會越請傾向投下反對票，反之亦然。

7.結論:

我們從上面做的各種相關性來看，其實發現就第十三案東奧正名來說，其實並沒有任何的因素是有明顯相關，必須要到使用三種因素，才勉強可以算是相關，不管是看regression 的r2或是看knn與decision tree的分類準確度都是如此，而其實全部特徵一起看也有一定的相關性，表示我們選的這七種特徵與算是與公投結果有一定的連結，並不是亂選的。而在第十案的婚姻平權來說，其實就讓我們滿訝異，我們原本想說年齡可能是有滿大的相關性，但是其實沒有，反而是教育程度的相關性出乎我們意料之外，不管是看趨勢圖或是r2值都是，且在同時使用三種因素的情形下，其實也都有不錯的相關性。我們可以說是以regression的r2為主，decision tree與knn的分類準確性為輔，來看待各種特徵，我們其實算是找到人驚訝也算成功的結論，算是完成了我們一開始動機想做的事情，而我們其實在做的事情都是客觀的分析數據，按照數據得到結論，但是究竟在現實中是甚麼原因會造成這些結論，其實我們做的結論是可以給一些社會研究或是政治探討來使用，我們也期待這件是發生。

貢獻度:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 李佳任 0510023 | 劉孟寰 0510022 | 方鈺豪 0510020 | 袁鈺勳 0510002 | 陳以嬿 0086043 |
| 22.5% | 22.5% | 22.5% | 22.5% | 10% |