# 程式設計與資料科學導論

# 期末專案-裁判書預測系統

（以酒駕犯罪進行分析探討）

組名:組名
成員:黃冠維(組長)、黃振瑋、李承翰、巫秉融

# 目錄

# 壹、前言

在當前的社會中，即便我國的法律已實行將近百年，但人民與法律的隔閡仍持續存在。常常可以在新聞媒體上看到「恐龍法官」、「殺一兩個人不會死」等等聳動的標題。對此，以民眾的角度而言，會認為法律不貼近社會現實，而法官也都缺乏社會歷練；但以法律人的角度而言，卻會認為大多的判決都是要經過嚴密的法律層面上的邏輯分析，才會得出判決結果的，考量的因素並非如同大眾所想的那麼單純。針對這樣的隔閡，我們希望可以做一項預測判決的專案。除了分析出法官在酒駕案件中，量刑的考量大多為何?影響的比重為何等外，可以讓吾人根據案件事實，輸入相對應的條件來取得模型預測刑度區間的結果。

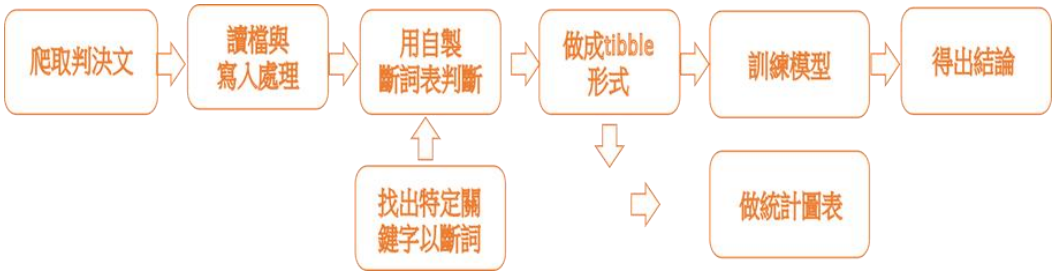
至於選取酒駕案件作為預測分析的原因為何，我們認為，首先，在觀察各類型的案件後，我們發現酒駕案件的事實相對單純，法官的論述方式、量刑考量也大多相似；其次，酒駕案件的案件數也十分充足，至民國 88 年以來，合計有近 20 萬篇的判決文，這對於需要大量 data 訓練的模型而言，非常友善；再者，酒駕案件也是當今社會最為關心的一項法律議題之一，以此作為分析，不僅貼近社會的需要，也因為相關的討論較多，在資源的查找上也會比較容易。

因此，我們選擇了酒駕案件來作為本專案的研究主題，若模型訓練順利，或許也可將這段期間所習得的技術應用在其他更為複雜的案件類型當中。

## 目標

1.了解酒駕犯罪中，影響法官判決的因素有哪些(權重為何?)
2.建立預測模型(Decision Tree)，在給定一定條件下，預測其判決結果
3.做出因素間關係的統計資料並進行分析與探討

# 貳、程式架構




## 一、爬蟲

1.考量到分析上的準確度與效率，我們搜尋的條件上須考量以下因素[1]

(1) 為避免修法造成適用刑度上的落差，故我們只取近兩年的判決(適用最新一次修法刑度)。

(2) 由於系統查詢結果包含裁定與判決，但我們需要的只有「判決」，故只取判決。

(3) 由於三級三審制度，使得同一案件可能會不只有一篇判決，造成誤差，故我們統一取地方法院判決。

(4) 考量資料處理時間不宜太久，故我們只取 108、109 年 5~12 月，每月 2 篇 500 篇的判決，共 8000 篇。

2.動態爬蟲

(1) 將上述網址儲存至 year_jud_list，並建立第一個迴圈(外層)依序(共 16 個網址)進行 navigate()。

(2) 使用 findElement()及 getElementText()抓取判決文內容，並儲存至 jud

(3) 將 jud 做空白處理，去除掉半形符號、全形符號、換行符號後存入 year_jud_list。

(4) 使用 findElement()及 clickElement()，進入到下一篇判決。

(5) 建立迴圈(內層)，重複上述(2)~(4)的動作，直到抓完 500 篇判決。

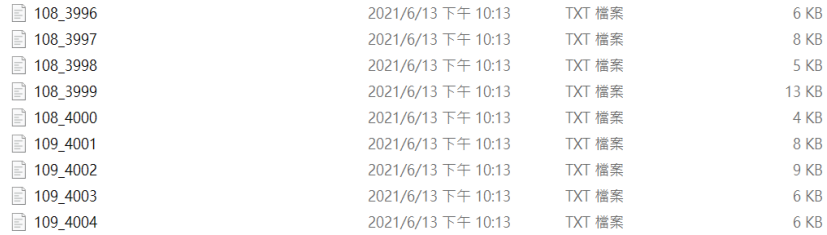
(6) 在兩層迴圈搭配下，成功抓取共 8000 篇判決。

[1] 由於這些條件的設定頗為複雜，在該網站有提供這樣的搜索功能下，我們選擇使用其搜索功能依序先儲存每月第一篇判決文的網址到程式碼中，再藉由 Rseleium 一一進入進行爬取

3.檔案處理

(1) 將上述儲存 8000 則判決的 year_jud_list，依序寫檔，並依據年度進行編號

(2) 如圖:




(3)使用時再將檔案讀入即可

## 二、資料處理

1.函數介紹(將切割好的判決文傳進函數取得我們要的資訊前，先定義以下 11 個函數)

| 函數 | 功能 | 演算法 |
|---|---|---|
| Court_f() | 判斷裁判法院 | 1.使用 grep()找出含有「地方法院」的詞的位置<br>2.將該位置 -1 取得地點位置，以找出該地點為何<br>3.將該地點後面加上『「地方法院」後回傳 |
| Date_f() | 判斷裁判日期 | 1. 使用 grep()找出含有「裁判日期民國」的詞的位置<br>2. 使用 grep()找出含有「裁判案由」的詞的位置<br>3.取出該二位置之間[2]所有的元素，並將各元素合併起來，成一詞<br>4.回傳該詞 |

[2] 因觀察判決文可以發現，這兩個詞之間的文字即為裁判日期(去除民國二字)

| | | |
|---|---|---|
| Result_f() | 判斷判決刑度 | 1. 使用 grep()找出含有「處有期徒刑」的詞的位置<br>2.使用 substr()取出該詞第 6、7 個元素(如:貳月、捌月等)[3]<br>3.回傳該刑度 |
| Result_level_f() | 判斷該刑度的區間 | 1.先傳入 Result_f()取得刑度<br>2.建立 if else 判斷式，回傳刑度區間<br>短:2 個月以下<br>中:3 個月~6 個月<br>長:大於 6 個月 |
| Alcohol_f() | 判斷酒精值 | 1.使用 grep()找出含有「毫克」的詞的位置<br>2.將該位置 -1 取得酒精值位置<br>3.回傳酒精值<br><br>*找不到「毫克」即回傳 NA[4] |
| Alcohol_level_f() | 判斷酒精值的區間 | 1.傳入 Alcohol_f()，取得酒精值<br>2. 建立 if else 判斷式，回傳刑度區間<br>low:0.25~0.54<br>medium:0.55~0.84<br>high: >=0.85 |
| Diploma_f() | 判斷學歷高低 | 1.先依照學歷 low、 |

[3] 惟此處出現瓶頸，1.大寫「三」判決書中可能出現成「參」或「叁」，但後者 R 顯示不出來，目前還想不到方法抓(但因後者比例較少，故影響不大)；2.此方法只能抓到刑度為兩個字的，若表示成「壹年伍月」即抓不到，雖然酒駕案件一年以上的判決極少，但實際對模型的影響尚無法得知

[4] 此處因發現，極少數判決是用「血液酒精濃度」單位為「MG/DI」，而非「呼氣酒精濃度」單單位為「毫克」

| | | |
|---|---|---|
| | | medium、high，分別建立對應的詞彙<br>2.建立 if else 判斷式，若判決文中出現該詞彙，則依此回傳學歷高低<br><br>*無此詞彙則回傳 NA |
| Economy_f() | 判斷經濟情況 | 1.先依照經濟情況 poor、medium、rich，分別建立對應的詞彙<br>2. 建立 if else 判斷式，若判決文中出現該詞彙，則依此回傳經濟情況<br><br>*無此詞彙則回傳 NA |
| Again_f() | 判斷是否累犯 | 1. 建立 if else 判斷式，判斷判決文中是否有出現「累犯」一詞。 |
| Accident_f() | 判斷是否肇事 | 1.依照是否肇事，分別建立相對應的詞彙[5]<br>2. 建立 if else 判斷式，若判決文中出現該詞彙，則依此回傳是否肇事<br><br>*無此詞彙則回傳 NA |
| Attitude_f() | 判斷犯後態度 | 1. 依照犯後態度好與不好，分別建立相對應的詞彙[6]<br>2. 建立 if else 判斷式，若判決文中出現該 |

[5] 此處為先參考酒駕判決使用的詞彙後，人工鍵入肇事與否會用到的詞語/句，惟因語意的表達方式多元，判斷結果多數皆為 NA 值

[6] 此處為先參考酒駕判決使用的詞彙後，人工鍵入肇事與否會用到的詞語/句，惟因語意的表達方式多元，判斷結果多數皆為 NA 值

| | | |
|---|---|---|
| | | 詞彙，則依此回傳犯後態度好與不好<br><br>*無此詞彙則回傳 NA[7] |

2. 資料擷取

(1) 使用 for 迴圈，將判決文傳入 segment()，進行切割後，再傳入上述函數[8]

(2) 將各函數回傳的值，依序存入不同的 list

(3) 將各 list 轉成向量後，建立 jud_df(tibble 形式)

(4) 如圖:

| | jud_id | court_place | date | result | result_level | alcohol | alcohol_level | diploma | economy | again | accident | attitude |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6964 | 6964 | 彰化地方法院 | 109年10月21日 | 貳月 | short | 0.90 | high | medium | medium | No | NA | good |
| 6965 | 6965 | 臺北地方法院 | 109年10月21日 | 肆月 | medium | NA | NA | medium | medium | No | No | good |
| 6966 | 6966 | 臺北地方法院 | 109年10月21日 | 貳月 | short | 0.57 | medium | medium | medium | No | NA | good |
| 6967 | 6967 | 宜蘭地方法院 | 109年10月21日 | 參月 | medium | 0.53 | low | medium | medium | No | NA | good |
| 6968 | 6968 | 臺北地方法院 | 109年10月21日 | 參月 | medium | NA | NA | medium | medium | No | No | NA |
| 6969 | 6969 | 彰化地方法院 | 109年10月21日 | 捌月 | long | 0.55 | medium | NA | NA | Yes | NA | good |
| 6970 | 6970 | 臺中地方法院 | 109年10月21日 | 貳月 | short | 0.53 | low | NA | medium | No | NA | good |
| 6971 | 6971 | 新竹地方法院 | 109年10月20日 | 肆月 | medium | NA | NA | NA | NA | Yes | NA | NA |
| 6972 | 6972 | 桃園地方法院 | 109年10月20日 | 貳月 | short | 0.62 | medium | low | medium | No | No | good |
| 6973 | 6973 | 桃園地方法院 | 109年10月20日 | 貳月 | short | 0.30 | low | low | medium | No | No | good |
| 6974 | 6974 | 桃園地方法院 | 109年10月20日 | 貳月 | short | 0.96 | high | low | poor | No | No | good |
| 6975 | 6975 | 桃園地方法院 | 109年10月20日 | 參月 | medium | 0.27 | low | low | medium | Yes | No | good |

Showing 6,963 to 6,976 of 8,000 entries, 12 total columns

## 三、模型訓練(Decision Tree)

1. 前置處理

(1) 使用 dplyr 套件的 filter()，將 jud_df 中任何有 NA 值的 Row 過濾掉，並整理成 tidytest(tibble 形式)

(2) 將 tidytest 隨機排序後，去除掉無關模型訓練的變數(編號、裁判法院、裁判日期、刑期[9]、酒精濃度區間)，並整理成 tidyenter。如下:

[7] 此處雖無上述問題，惟經觀察法官僅會在被告犯後態度好時才會於判決文中提及(故沒有 bad 值)，造成本因素判斷失真

[8] 為求正確切割，不同的函數我們都有設計不同的切割依據，故在傳入每個函數前，都要先做一次切割

[9] 因本預測模型是欲預測出「刑度區間」而非「刑期」

| | result_level | alcohol | diploma | economy | again | accident | attitude |
|---|---|---|---|---|---|---|---|
| 843 | medium | 0.89 | low | poor | Yes | No | good |
| 844 | short | 0.89 | low | medium | No | No | good |
| 845 | short | 0.69 | medium | medium | No | No | good |
| 846 | medium | 0.56 | medium | medium | No | No | good |
| 847 | medium | 0.38 | medium | medium | No | No | good |
| 848 | short | 0.46 | low | poor | No | No | good |
| 849 | medium | 0.69 | low | poor | Yes | No | good |
| 850 | medium | 0.56 | low | medium | Yes | No | good |
| 851 | medium | 0.50 | low | medium | No | No | good |
| 852 | medium | 0.87 | low | medium | No | No | good |
| 853 | medium | 0.35 | medium | medium | No | No | good |
| 854 | medium | 0.59 | low | poor | No | No | good |
| 855 | short | 1.00 | high | medium | No | No | good |

Showing 843 to 855 of 1,171 entries, 7 total columns

[10]

(3) 將 tidyenter 依刑度高低，分為三個 Tibble 後，分別再依 7:3 的比例做分割，最後組成 trainingdata 和 testdata

2. 模型訓練

(1) 使用 rpart 套件中的 rpart()函數，將 trainingdata 傳入進行訓練

(2) 使用 rpart 套件中的 predict()函數，將 testdata 傳入，取得測試結果

## 四、製作圖表

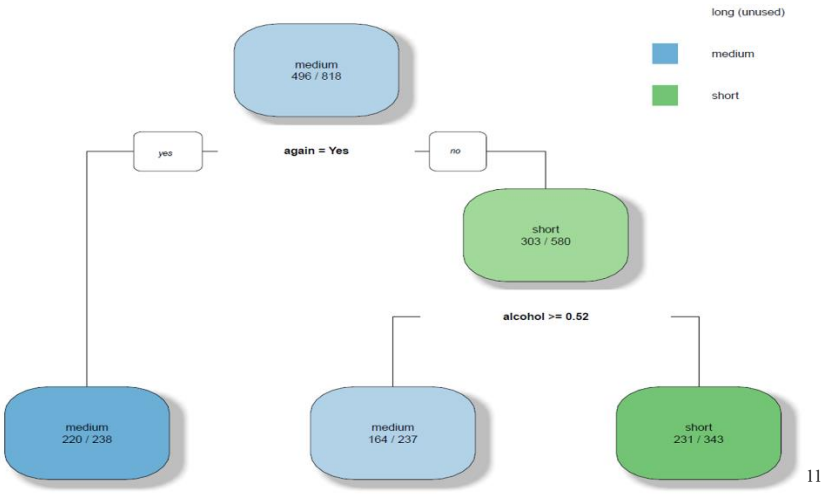
1. 使用 rpart.plot 繪製決策樹圖

2. 使用 ggplot2 繪製各變數間的關係圖

[10] 注意:有效數據從 8000 篇下降為 1171 篇，主要原因為學歷、經濟狀況及是否肇事出現太多 NA 值，可見資料處理方面，需要再多下點功夫(不過一部分的原因也是因為這三項並不是判決總是會出現的項目)

# 參、成果展現

## 一、模型結果



[11]

## 二、準確度≈ 72%

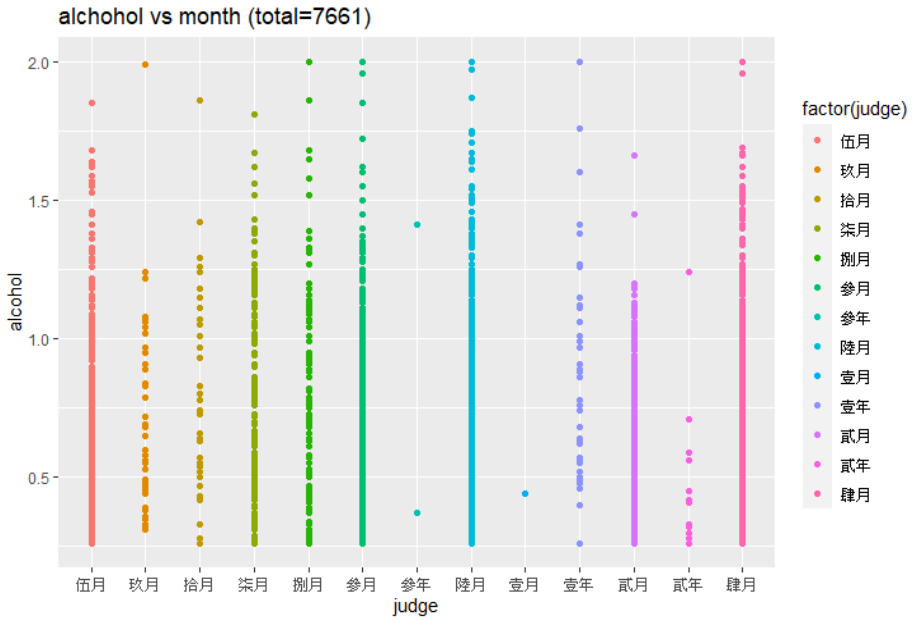
| Predict\Data | long | medium | short |
|---|---|---|---|
| long | 0 | 0 | 0 |
| medium | 4 | 160 | 40 |
| short | 0 | 54 | 95 |

(計算式:(0 + 160 + 95) / 353 ≈ 0.72234 )

[11] 此圖可看出，累犯與否及酒精值濃度的高低是影響刑度區間的重要因素。而高刑度則因為樣本數不足，被模型歸類「unused」。這部分主要是因大於陸月刑度的酒駕犯罪本身就很少，而上述資料處理的問題(註 3)又加劇了這種現象，使得樣本數不足，尚待解決。

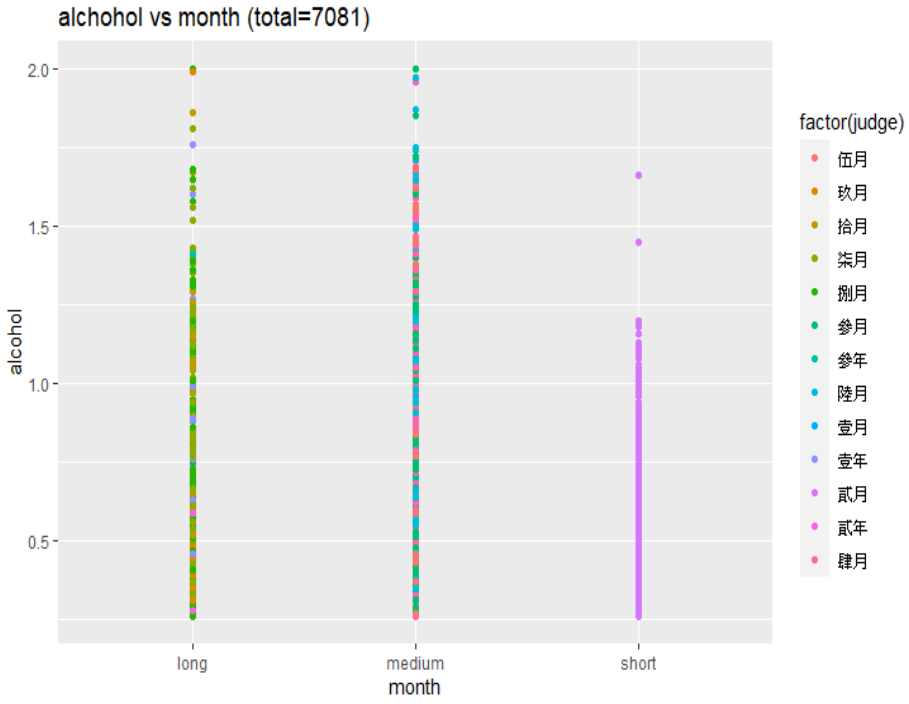
## 三、圖表分析

1. 酒精濃度與刑罰輕重的相關性




（圖一）酒精濃度與實際刑度圖

（每個點代表刑責(x)相對於酒精濃度(y)）



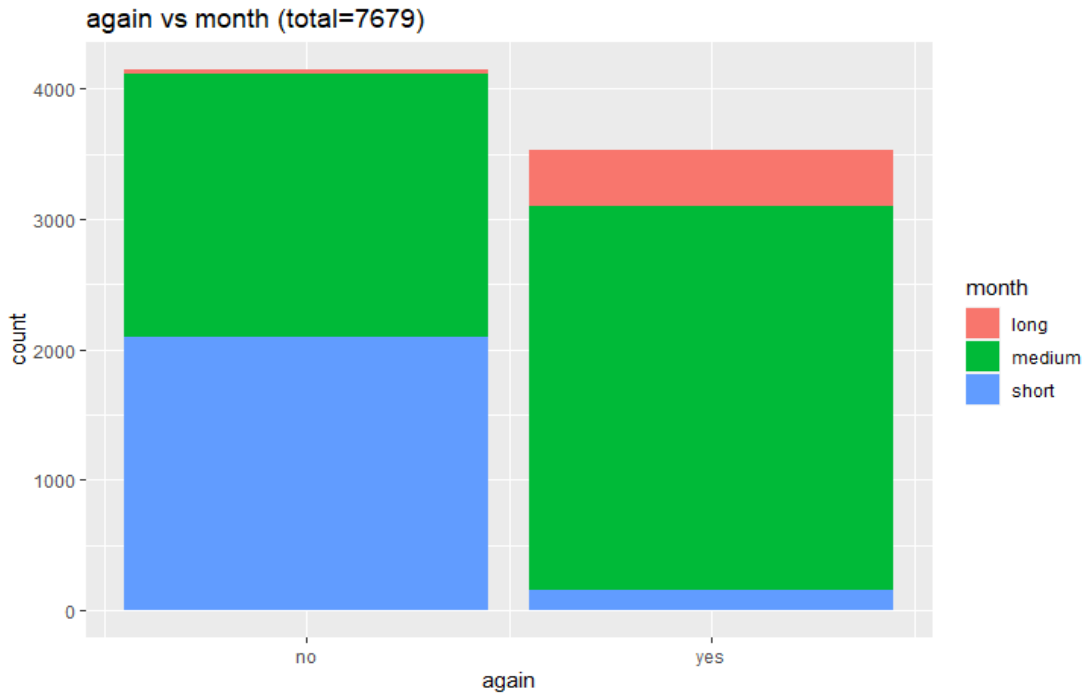

（圖二）酒精濃度與相對刑度圖

觀察

(1) 刑責較輕的有較大機會酒精濃度較低，符合模型預測

(2) 刑責較重的酒精濃度並不一定比刑責輕的還高，可能是由於其他因素(肇事、累犯等)影響，在模型判讀上會有誤差

2. 累犯與刑罰輕重的相關性



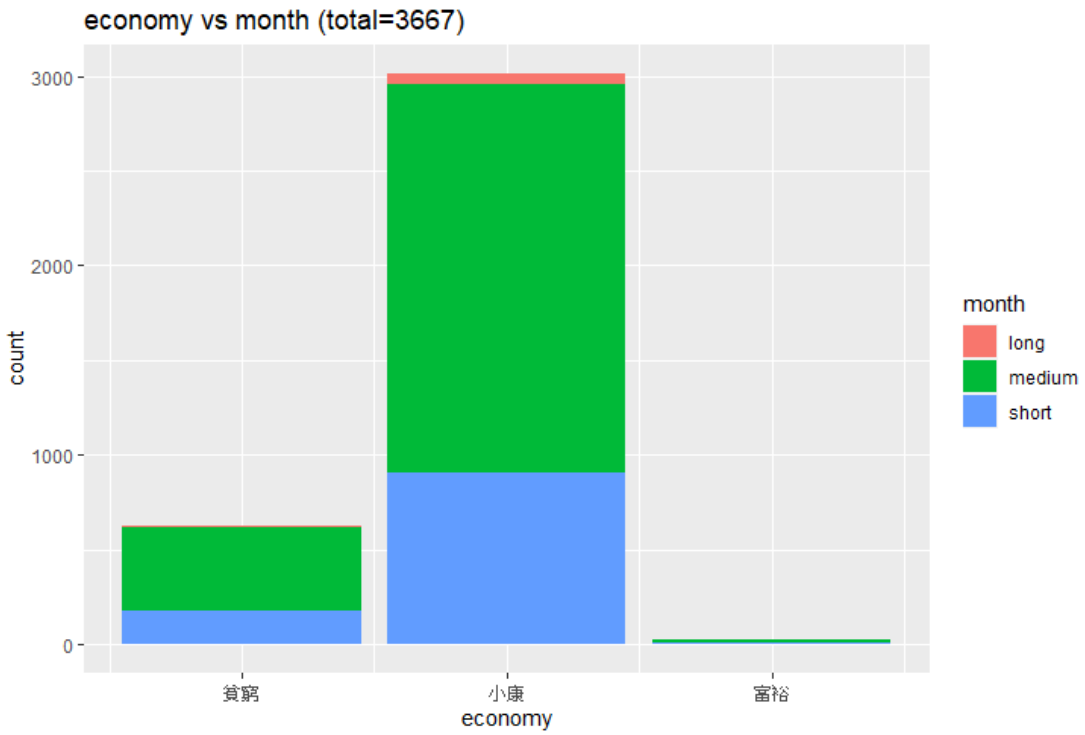

(圖三)累犯與相對刑度累計圖

觀察:

(1) 累犯的案件數略低於初犯

(2) 累犯中，刑度較高的比例較多，符合模型預測

3. 經濟狀況與刑罰輕重的相關性



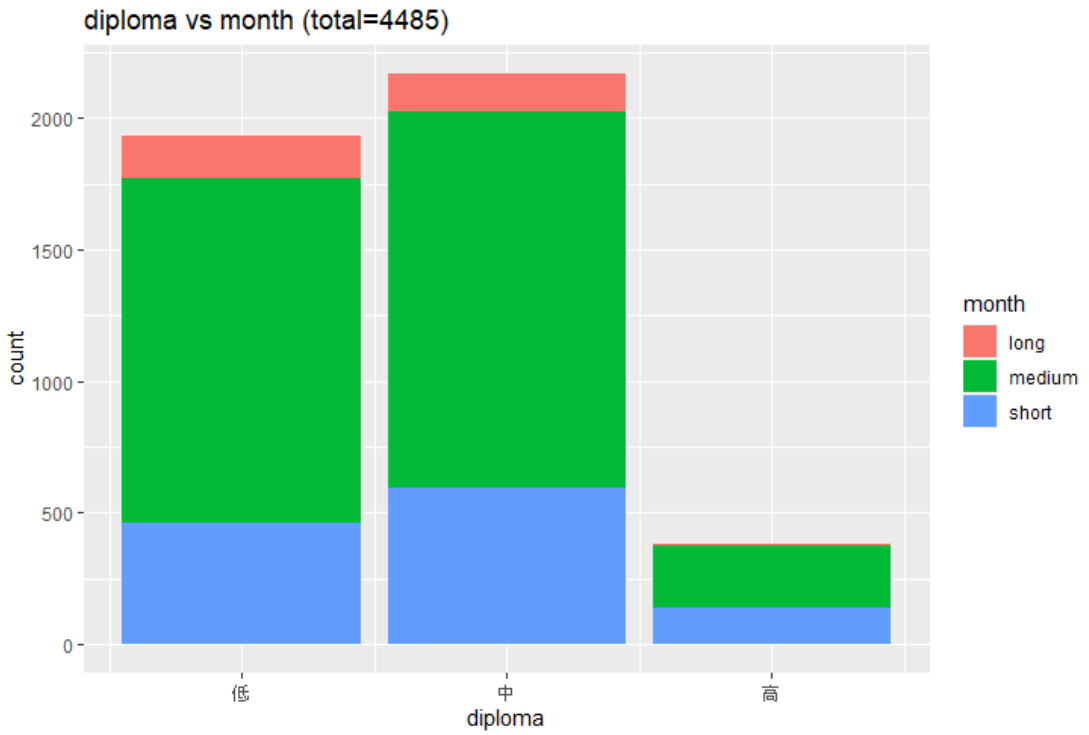

（圖四）經濟狀況與相對刑度統計圖

觀察

（1）數據顯示家境小康的占數據很大一部分，雖然富裕的樣本可能並沒有參考性，但三千多份資料佔據整體數據約四成的分量，顯示酒駕事件的家庭很可能一大部分來自小康家庭。[12]

4. 學歷與刑罰輕重的相關性

[12] 但此處具數據上的誤差，因法院通常只有被告經濟普通或困難時才會提及，以至於較富裕的被告在前面的資料處理及被列為 NA 值而被排除在外了。

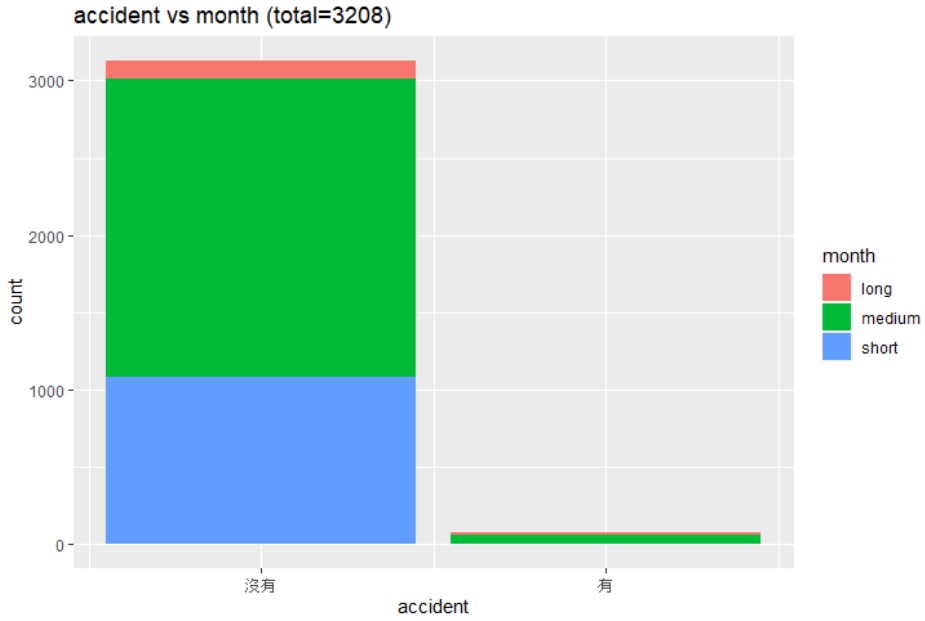



(圖五)學歷與相對刑度統計圖

觀察:

(1) 由於可判定的詞彙較為精確(高中、國中等),因此數據相對有參考性,可發現高學歷的案件數較其他的群數量較少(也有可能是法官鮮少提及高學歷的部分),但撇除掉此前提的情況(高學歷的數量並沒有少到沒有參考性),整體來說高學歷的刑罰較低。(可以由綠藍紅的的比例來看)

5. 肇事與刑罰輕重的相關性

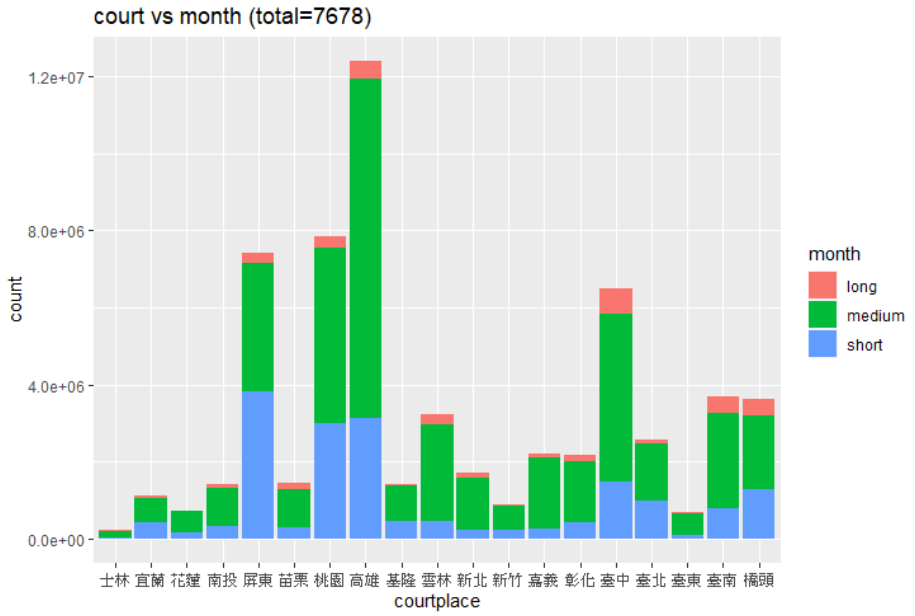



(圖六)肇事與相對刑度統計圖

觀察:

(1) 由於有肇事的詞彙並不容易抓取判定，造成沒肇事的比例較為多，事實上，法官在描述肇事的部分有可能會依據當時的情況陳述事實，而非簡單的「幸無肇事」來帶過。此問題造成判讀成功的比例較為懸殊，顯不具參考性。

6. 地方法院處裡的酒駕案件數




(圖七)地方法院處理案件統計圖

觀察:

(1) 由整體的數據比例可看出高雄的酒駕案件數站採樣數據的一大比例(高雄與橋頭)。桃園跟屏東為其次。台北以及新北的人口位居台灣第一及第四，但酒駕案件數比較低，似乎不能用人口數單向推測酒駕案件數的多寡。當地的風俗名情與地方治安為需要考量的部分。

## 四、小結

1. 酒精濃度對於刑度有重要的影響。亦即，案件要被判為輕刑者，濃度大多需要較低(輕型中僅有兩例為濃度 1.25 毫克以上)，惟酒精濃度低，並不表示刑度較低，仍須考量其他因素。

2. 累犯與否對於刑度有重要的影響。亦即，欲被判為輕刑者，大多必須為非累犯(只有不到 3%的累犯者會被判輕刑)。反之若為累犯，則大多會被判為中刑或重刑。

3. 其他因素如:經濟情況、學歷、犯後態度、肇事與否對於刑度並無明顯關聯，惟「肇事與否」可能是因資料擷取時出了問題，使得樣本數不足。

4. 透過此模型，我們可以進行判決預測。方法為:
   (1) 由當事人輸入其案件的特徵(如:被測得的酒精值、是否累犯等)
   (2) 將該特徵整理成 tibble 形式並丟入模型
   (3) 得出模型預測的刑度區間

# 肆、結論/未來展望

在這次的專案中，「酒精濃度」與「是否累犯」兩變項，果然是影響法官量刑的重要因素，符合當初的設想。但原本亦認為「肇事與否」也會對量刑有重大的影響，但是因為前述資料擷取效果不佳的問題，使得結果不如預期。我們認為可能的嘗試方法為，利用向量表徵的作法，來讓相似的語句皆能被成功偵測，而非透過人工方式一一輸入，這樣或許就能降低語言表達具多異性對專案的影響了，

此外，我們查找相關研究文獻時也有發現，目前的判決預測多是透過人工標記判決文的方式，標記出「類型特徵」及「理據特徵」[13]，前者如:雙方年齡、國籍、判決結果、身分等；後者則為有利與不利雙方的語句，來產生Data。此舉確實對增強判決預測的準確度有極大的助益(因機器尚較無法準確偵測的語言表達的多異性)，但是卻須要耗費大量的人力、物力資源。而此專案即是在「準確度」與「所需成本」間所做出的衡量結果。如前言中所述，酒駕的判決大多案件事實單純，證據清楚[14]，且論理結構也相近、量刑因素相似、判刑結果較為集中、法條多僅涉及刑法第一百八十三條之三第一項[15]等。基此，才創造「酒駕犯罪」可以使用「非人工」的方式進行文句偵測的條件。

最後，在這項專案，我們也學習到了很多程式應用上所需的技能。如在爬取判決文時，遇到透過 Rvest 無法抓取的情形，使得我們必須去了解 Rselenium 這個套件、並下載相關軟件(如: chromedriver、jdk、Rwebdrive)並了解其運作方式。而在網頁標籤的查找下，我們也費了許多功夫才成功抓到判決文(如:

---

[13] 未來可以往此處邁進

[14] 多不會有發生抗辯的問題，被告大多會直接認罪

[15] 該條雖依照不同條件有不同的刑度區間適用，但酒駕犯罪大多皆為適用第一項，致重傷或致死的情況即為少見，下為刑法第 185-3 條

「駕駛動力交通工具而有下列情形之一者，處二年以下有期徒刑，得併科二十萬元以下罰金：

一、吐氣所含酒精濃度達每公升零點二五毫克或血液中酒精濃度達百分之零點零五以上。

二、有前款以外之其他情事足認服用酒類或其他相類之物，致不能安全駕駛。

三、服用毒品、麻醉藥品或其他相類之物，致不能安全駕駛。

因而致人於死者，處三年以上十年以下有期徒刑；致重傷者，處一年以上七年以下有期徒刑。

曾犯本條或陸海空軍刑法第五十四條之罪，經有罪判決確定或經緩起訴處分確定，於五年內再犯第一項之罪因而致人於死者，處無期徒刑或五年以上有期徒刑；致重傷者，處三年以上十年以下有期徒刑。」

遇到巢狀頁面的問題[16])。此外，在決策樹模型上我們也透過實際操作，更了解老師在課程後期所講述的機器學習概念，對未來的延伸學習頗有打開一扇窗之感。感謝這次的專案讓我們更了解 R 語言的程式邏輯以及運用，相信有了這次程式能力突飛猛進的經驗，在未來，我們在程式語言的學習上，將能更上層樓，邁向更深、更廣的學習。

# 伍、分工

| 負責項目 | 組員 |
| --- | --- |
| 專案發想 | 黃冠維(組長)、黃振瑋、李承翰、巫秉融 |
| 爬蟲 | 黃冠維、李承翰、巫秉融 |
| 資料處理 | 李承翰、巫秉融 |
| 模型訓練 | 李承翰、巫秉融 |
| 成果展現(含圖表、投影片製作) | 李承翰、巫秉融 |
| 口頭報告 | 李承翰、黃冠維(6/10)<br>李承翰、巫秉融(6/17) |
| 書面報告、繳交 | 李承翰、巫秉融 |

[17]

# 柒、專案連結

https://github.com/rlads2021/project-KevinLee1335

[16] 雖然後來是用「人工」加「非人工」的方式共同處理，實際上沒有解決巢狀頁面爬取的問題

[17] 本組在分工上出了些問題，主要是因前期討論不足，專案目標鬆散所致，使得無進行分工，呈現兄弟各自努力，但不知他人進度的情況。也使得 6/10 進度報告前數天才完成了「爬取判決文」這項步驟