HW2

0513250 邱郁雯

1. 清洗Accident.csv的資料
2. A1/A2資料合併成A2

將A1/A2資料合併成一類(A2)，A3另成一類分析。

1. 資料不列入

不將編號、天候名稱、光線名稱、路面狀況名稱、當事者性別名稱、車種名稱、保護裝備名稱及飲酒情形名稱列入資料分析內容，因為我認為編號對於分析結果應沒有幫助，只是普通的數字，其他天候、光線、路面狀況等名稱，因為他們有各自對應的代碼，所以不將這些說明列入分析中。

1. 資料內容改動

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 天氣代碼 | 1 | 2 | 4 | 6 | 7 | 8 |
| 天氣名稱 | 暴雨 | 強風 | 霧或煙 | 雨 | 陰 | 晴 |

因accident內的天候代碼與天候名稱對不上，例如天候代碼8對應的名稱就有「晴」與「暴雨」，無法做惡劣天氣的分類，因此我將每個代碼都只有對應一個天氣名稱。

1. 遺失資料處理

* 一開始匯入遺失的資料筆數：
* 將空白的資料以比數最多的資料取代：

找出比數最多者、將資料取代後，確認沒有遺漏的資料。

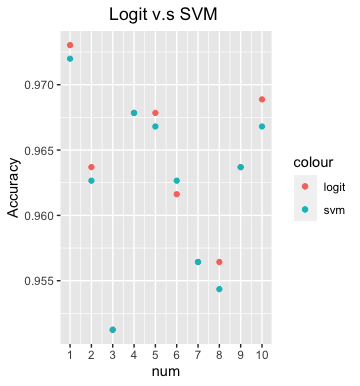
* 車種代碼為factor，所以使用dummy variable分析：

1. 使用SVM方法及Logistics方法進行10次預測，預測目標為「事故類別」。

* 進行10次的Logistics方法與SVM方法預測事故類別：

在logistics中，我認為事故類別裡A3的發生次數會較多，所以機率應該較A1, A2高，所以當機率大於0.5則分類為A3。

* Logistics和SVM的準確度：

這兩個方法預測事故種類的準確度都高達95%以上，在圖中基本上logit的準確度都比SVM高一些，而在第3,4,7,9筆資料中，SVM和logit的準確度一樣，所以在圖片中只有一個點。