Report

目的Abstract

此次主要目標為「依車輛行車事故鑑定，來預測車禍案件中法院判定為過失案件。」，資料設計的分析主要對象為依交通部所發布「車輛行車事故鑑定及覆議作業辦法」進行鑑定。資料設計的分析主要徵對台灣機車的交通事故案例。

交通過失案件是指在道路上發生的交通事故，其中一方或多方被認定為有過失，即他們在事故中未能遵守交通法規或行車規則，導致了事故的發生。在非過失的情況下，通常不需要對非過失方進行處罰或賠償，因為他們被認定為已經盡到了駕駛的責任，並遵守了相關法規。不過，對於過失方，根據法律可能需要負擔相應的法律責任，包括支付賠償金、罰款或其他處罰。

特徵設定 Setting

Input特徵設定

1. 事故時間(0~24)：記錄事故發生的具體時間，以小時為單位。
2. 降水量：以毫米為單位的數字。
3. 風速：表示當天風速，以 公里/小時 為單位的數字。
4. 駕駛速度：記錄每輛參與事故的車輛的速度，以 公里/小時 為單位的數字。
5. 駕駛人酒精濃度：0.00（無酒精）- 0.12（高濃度酒精）
6. 違反交通信號(0, 1)：0表示沒有，1表示有，例如闖紅燈、迴轉禁止、未遵從行人穿越標示等。
7. 路寬：必須大於等於1.5，以公尺為單位的數字。
8. 路面狀況(0-5)：0表示平坦，5表示崎嶇。
9. 肇事者精神狀況(0-10)：0表示身心異常，10表示健全。
10. 騎車技巧熟練(0-10)：0表示新手，10表示技巧熟練。
11. 交通事故類別（1: A1, 2: A2, 3: A3）：根據交通事故處理規範。（A1類：造成人員當場或二十四小時內死亡之交通事故、A2類：造成人員受傷或超過二十四小時死亡之交

通事故、A3類：僅有財物損失之交通事故。）

Output 特徵設定

- 過失(1)：有
- 非過失(0)：無

Absolute right rule

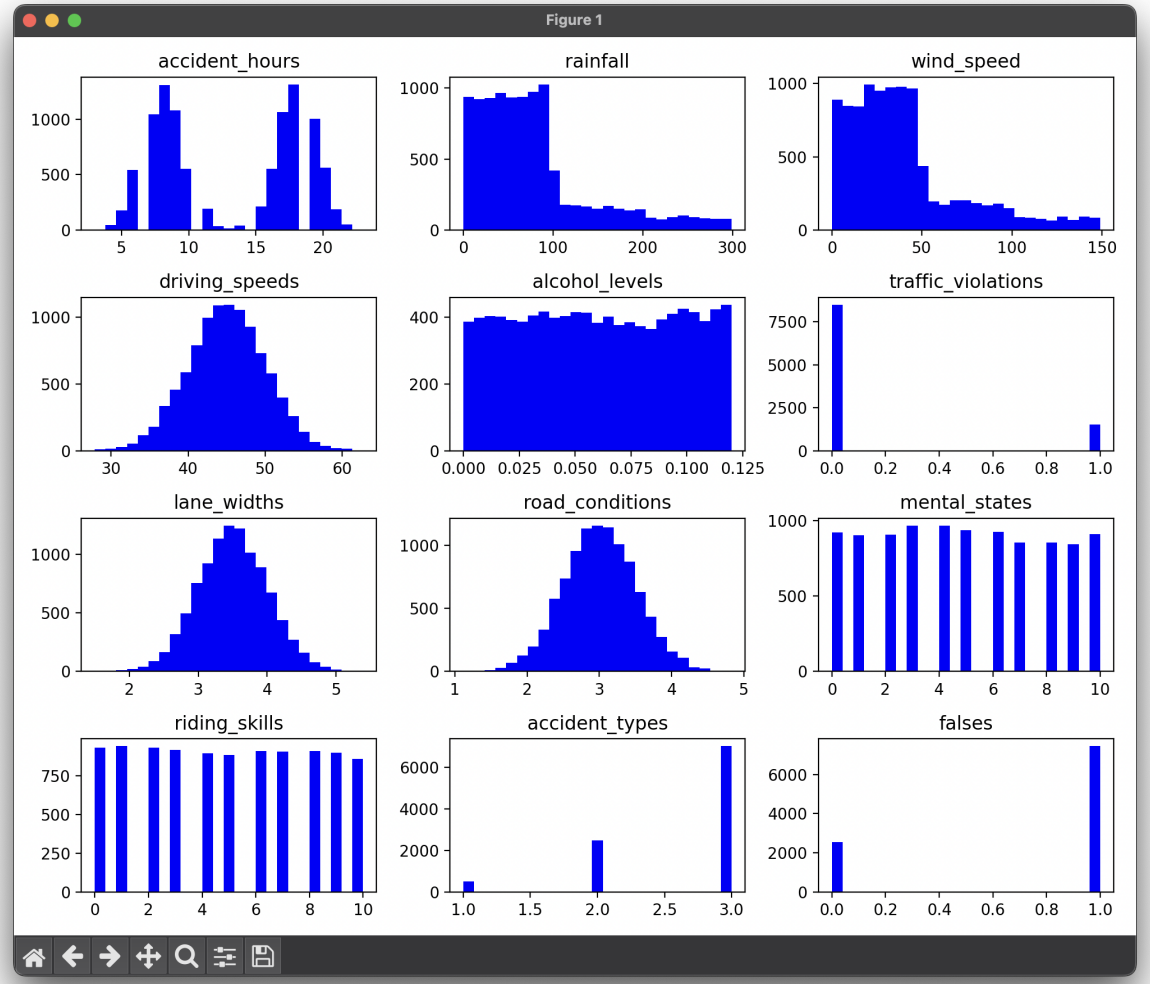
1. 嚴重交通事故類別規則：
    a. 如果交通事故類別為A1（造成人員當場或二十四小時內死亡），則此類事故通常會被判定為過失案件。
2. 違反交通規則：如果駕駛人違反了交通規則（如闖紅燈、酒駕），這很高機率會被判定為過失案件。
    a. 駕駛人違反了交通信號（如闖紅燈、壓線）或者是 駕駛人酒精濃度超過0.08（一般被視為酒駕的法定標準）
    b. 駕駛速度遠高於該路段的法定速度限制（超過限速=路寬40）
3. 惡劣天氣條件規則：這個組合考慮到凌晨時段視野受限，再加上高降水量事故風險可能增加，這可能被視為過失駕駛。
    a. 發生在凌晨2點到5點之間
    b. 降水量大於50毫米
4. 精神狀態與狹窄道路結合規則：這種情況下駕駛者可能處於酒後狀態，同時道路條件也較差。這樣的組合增加了事故發生的風險。
    a. 駕駛人的酒駕 或是 精神狀況不佳（駕駛的精神狀況超過4）
    b. 路寬小於2公尺
5. 高速駕駛與路面不平結合規則：這種情況下高速駕駛和不平坦的路面可能導致車輛控制困難。這樣的組合是事故高風險的一個指標。
    a. 駕駛速度超過70公里/小時
    b. 路面狀況評分大於4（表示路面相對崎嶇）

Setting 數據集相關設定

- 交通事故的時間是依照上下班時間的高斯分佈來產生，這個分佈模擬了在早上和晚上上下班高峰時間段發生事故的概率較高的情況。
- 降水量和風速為多項式分佈。根據台灣的氣候特點，在這個分佈中5-9月（梅雨季和夏季）的降水量占總降水量的70%，其餘月份占30%。同時，7-9月期間的降水量設定為100-299毫米範圍，以反映夏季可能出現的較高降水量，在 5-9 月日雨量超过 200 毫米的概率更高，如果生成的降水量数据小于 1000，则用 0-99 毫米的降水量填充。風速的分布，假設 5-9 月是颱風季節這段時間的風速可能會更高，定義為多項式分布的概率。
- 駕駛速度根據高斯分布參數生成（平均時速為45km/hr，標準差為5），數據集中的速度值被限制在0到120 km/hr 之間。
- 使用均勻分佈隨機生成了酒精濃度數據
- 生成違反交通信號的數據集，0的概率為85%，1的概率為15%
- 路寬根據高斯分佈生成，平均值為台灣單一車道寬度的設計標準3公尺，並且確保所有數據值均大於等於1.5公尺。
- 路面狀況以高斯分佈生成，平均值為3，標準差為0.5，並且將數據值限制在0到5的範圍內。
- 騎車技巧熟練度，值域範圍為0到10，以均勻分配隨機生成
- 交通事故類別，根據多項式分布的概率，概率分別調整為5%, 25%, 70%。

資料呈現 Data Appearance

經過以上之設定完成後，我們加以根據設定之特徵生成資料，共生成10000筆(policy_data.csv)。總共11項特徵之資料（最後一行「falses」即為此次分類目標），如下所示：



模型建立與評估 Model and Evaluation

本次作業一共選取了下列幾四種sklearn套件中的模型進行建模與比較，依序為：
決策樹、KNN、邏輯迴歸、高斯貝氏分類器，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），之後分別將資料帶入不同的模型，最終比較其模型分類之差別以找出其中關鍵因素。以下的兩張圖為輸出結果。

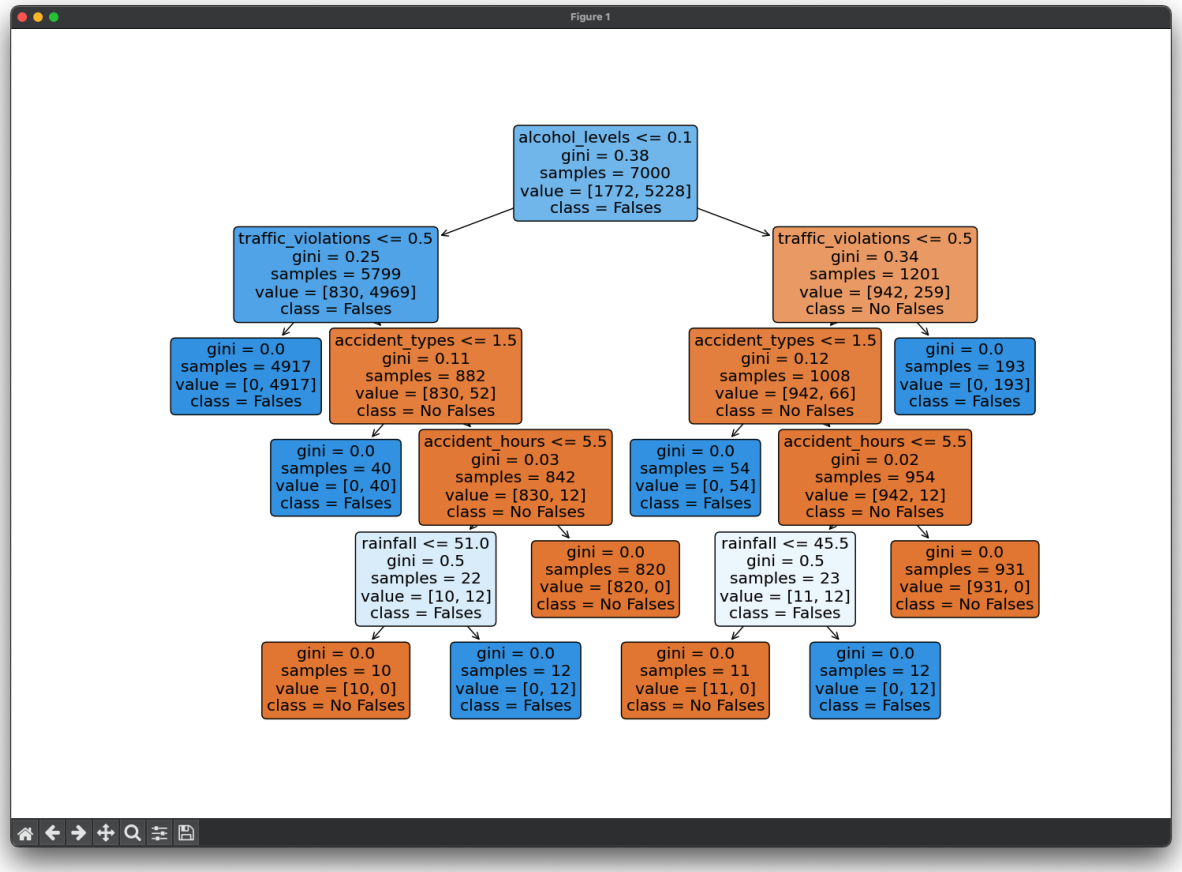
```
(project-env) zoelin@ZoeLin2023:~/Desktop/DataMiming-Project2(main⚡) » python3 main.py
Model: Decision Tree
Accuracy: 0.9997
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       767
           1       1.00      1.00      1.00      2233

    accuracy                           1.00      3000
   macro avg       1.00      1.00      1.00      3000
weighted avg       1.00      1.00      1.00      3000

--------------------------------------------------------------------------------
Model: K-Nearest Neighbors
Accuracy: 0.6943
Classification Report:
              precision    recall  f1-score   support

           0       0.27      0.11      0.16       767
           1       0.75      0.89      0.81      2233

    accuracy                           0.69      3000
   macro avg       0.51      0.50      0.49      3000
weighted avg       0.62      0.69      0.65      3000

--------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------
Model: Logistic Regression
Accuracy: 0.8267
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.44      0.57       767
           1       0.83      0.96      0.89      2233

    accuracy                           0.83      3000
   macro avg       0.81      0.70      0.73      3000
weighted avg       0.82      0.83      0.81      3000

--------------------------------------------------------------------------------
Model: Gaussian Naive Bayes
Accuracy: 0.8260
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.45      0.57       767
           1       0.83      0.96      0.89      2233

    accuracy                           0.83      3000
   macro avg       0.81      0.70      0.73      3000
weighted avg       0.82      0.83      0.81      3000

--------------------------------------------------------------------------------
```

決策樹 (Decision Tree)

準確率 (Accuracy): 99.97%

決策樹模型在測試集上幾乎完美地進行了預測。這可能意味著模型對數據進行了過度擬合，即它太過適應了訓練數據，以至於可能無法泛化到新的數據上。



K-最近鄰 (K-Nearest Neighbors, KNN)

準確率 (Accuracy): 69.43%

KNN模型的整體準確率較低，這可能意味著該模型沒有很好地捕捉到數據的特徵或者該數據集不適合使用KNN進行分類。

邏輯回歸 (Logistic Regression)

準確率 (Accuracy): 82.67%

邏輯回歸模型在預測非過失事故（標籤0）時表現良好，但在預測過失事故（標籤1）時的召回率低。這表明模型在區分過失和非過失事故方面存在一定的困難。

高斯貝氏分類器 (Gaussian Naive Bayes)

準確率 (Accuracy): 82.60%

高斯貝氏分類器的性能與邏輯回歸相似，這可能表明它在這組數據上受到了相似的限制。

結論 Conclusion

本次研究旨在參考台灣的交通狀況，透過各種特徵變數的模擬來預測交通事故中的過失案件。在理想狀態下，模型在訓練集和測試集上的表現應該相近。實驗過程中，不斷調整特徵變數，以期達成更自然的預測結果。實驗發現，決策樹模型的準確率異常高，接近100%，這可能指示數據過於簡單或模型發生了過度擬合。在實際應用中，這種情況似乎是非常罕見的。

而KNN模型的表現不佳可能意味著需要進一步的特徵工程或嘗試其他算法。邏輯回歸和高斯貝氏的表現相似，暗示數據特徵與目標變量的關係可能符合這些模型的基本假設。然而，這些模型在過失案例的召回率較低，顯示預測過失的能力有限。

透過多次實驗，我們了解到並非所有資料都適合同一模型。為了提高分析的準確性，必須考慮資料特徵是否符合模型假設，並根據模型在資料中的表現來進行多次測試和調整，以找出準確率低或過度擬合的原因。

此次實驗可能存在的不足和未考慮到的方面可能包括數據清洗和預處理做得不充分，可能會導致模型表現不佳。數據集也可能缺乏足夠的多樣性，無法充分捕捉到現實世界的複雜性。在模型評估上單一的性能指標（如準確率）可能不足以全面評估模型的性能，應使用更全面的指標（如ROC曲線、AUC值等）。