AI Capstone Project 1 111550056陳晉祿

**Topic: Analysis and Prediction of Traffic Accident Categories**

**Motivation:**  
Traffic accidents occur daily for a variety of reasons and are often unavoidable. Major incidents frequently make headlines, raising the question: how can we minimize the resulting harm? Traffic accidents involving casualties are classified into two categories: A1, where fatalities occur, and A2, where injuries are sustained without any deaths. By leveraging predictive models to rapidly forecast the potential category of an accident, emergency responders and healthcare professionals can more effectively allocate resources, ultimately reducing casualties and mitigating harm.

**Dataset Documentation: Taiwan Traffic Accident Data (A1 & A2)**

**1. Overview**

This dataset contains traffic accident records from Taiwan with information obtained from the [政府資料開放平台 (DATA.GOV.TW)](https://data.gov.tw). After process, it has been split into two main categories:

* **A1**: Fatal accidents (致死車禍)
* **A2**: Accidents involving only injuries (不含死亡之傷害事故)

Each row in the dataset represents a single accident event, with columns indicating various attributes such as date/time, weather conditions, road conditions, the types of vehicles involved, the age ranges of the people involved, and so on.

**2. External Source**

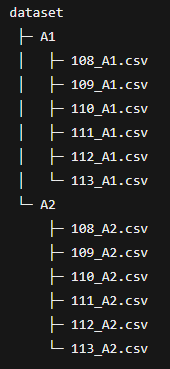
* **Source**: [政府資料開放平台-歷史交通事故資料](https://data.gov.tw/dataset/12197)
* **Original Format**: CSV files containing detailed accident reports (in Chinese), and each row represented a single person involved in an accident with columns such as date, time, categories… Therefore, if multiple people were involved in one accident, the original data would have multiple rows with identical accident date/time/location fields but differing personal or vehicle information.



**3. Data Composition and Amount**

**3.1 File Structure**

You will find multiple CSV files in two main folders:



* **A1** folder: Contains CSV files for A1 category (fatal accidents) across different years 108(2019) to 113(2022).
* **A2** folder: Contains CSV files for A2 category (injury-only) accidents for the same years.

**3.2 Rows and Columns**

* **Rows**: Each row corresponds to a single accident.
* **Columns**: After processing, the dataset has the following columns (all integer type after mapping):
  1. 發生日期 (accident date)
  2. 發生時間 (accident time)
  3. 天候名稱 (weather) – mapped to integer codes
  4. 光線名稱 (light condition) – mapped to integer codes
  5. 速限-第1當事者 (speed limit for the first party involved)
  6. 路面狀況-路面狀態名稱 (road surface condition) – mapped to integer codes
  7. 號誌-號誌種類名稱 (traffic signal type) – mapped to integer codes
  8. male (number of males involved)
  9. female (number of females involved)
  10. veh\_機車 (number of motorcycles involved)
  11. veh\_大客車 (number of large passenger vehicles involved)
  12. veh\_人 (number of pedestrians involved)
  13. veh\_大貨車 (number of large trucks involved)
  14. veh\_小客車(含客、貨兩用) (number of small passenger/cargo vehicles involved)
  15. veh\_慢車 (number of slow-moving vehicles involved)
  16. veh\_小貨車(含客、貨兩用) (number of small trucks, including passenger/cargo type)
  17. veh\_全聯結車 (number of full-trailer trucks)
  18. veh\_曳引車 (number of tractor trucks)
  19. veh\_半聯結車 (number of semi-trailer trucks)
  20. veh\_其他車 (number of other vehicles)
  21. veh\_特種車 (number of special-purpose vehicles)
  22. veh\_小貨車 (number of small trucks)
  23. veh\_軍車 (number of military vehicles)
  24. age\_1~10 (number of people aged 1–10)
  25. age\_11~20
  26. age\_21~30
  27. age\_31~40
  28. age\_41~50
  29. age\_51~60
  30. age\_61~70
  31. age\_71~80
  32. age\_81~90
  33. age\_91~100
  34. age\_101~110
  35. age\_111~120

**Note**: All columns are integers (int).

**3.3 Data Volume**

* Each A1 category CSV file (e.g., 108\_A1.csv, 109\_A1.csv, etc.) contains around 1600 records. And Each A2 category CSV file contains around 250000 records. The exact number depends on the number of accidents reported in that year.

**4. Data Collection Conditions**

1. **Time Frame**: Data covers the years 108 to 113 in the Minguo calendar (approximately 2019 to 2024 in the Gregorian calendar), though the actual availability of data may depend on the original government platform’s updates.
2. **Geographical Coverage**: All reported accidents occurred within Taiwan.
3. **Accident Severity**: Two categories:
   * A1 (fatal accidents)
   * A2 (injury-only accidents)
4. **Filtering Criteria**:
   * Only rows whose categorical columns match the predefined mappings (e.g., 天候名稱 in [晴, 陰, 雨, 霧或煙, 風, 風沙, 雪, 強風, 暴雨]) are retained.
   * Rows with invalid or missing data in the mapped columns are excluded.

**5. Data Processing and Tools Used**

1. **Original Data Download**
   * The original CSV files were downloaded from the [政府資料開放平台](https://data.gov.tw/dataset/12197).
   * The dataset included detailed text descriptions of road conditions, weather, signal types, and more.
2. **Mapping to Numeric Codes**
   * A Python function mapping(df) was applied to transform textual columns into integer codes:
3. **Software and Environment**
   * **Python** (Pandas) for data cleaning and mapping.
   * **CSV** format for storing processed data.
   * **Visual Studio Code** or similar IDE for code editing and data inspection.
4. **Hardware**
   * Standard desktop or laptop computer with enough RAM to handle large CSV files. No specialized hardware was required beyond typical development environments.

**6. Examples**

**6.1 Example Row (Processed Data)**

| **發生日期** | **發生時間** | **天候名稱** | **光線名稱** | **速限-第1當事者** | **路面狀況-路面狀態名稱** | **號誌-號誌種類名稱** | **male** | **female** | **veh\_機車** | **veh\_大客車** | **veh\_人** | **…** | **age\_21~30** | **age\_31~40** | **…** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20190101 | 12800 | 0 | 4 | 60 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | … | 0 | 0 | … |

* **天候名稱 = 0** → 晴
* **光線名稱 = 4** → 夜間(或隧道、地下道、涵洞)有照明
* **路面狀況-路面狀態名稱 = 0** → 乾燥
* **號誌-號誌種類名稱 = 0** → 無號誌

**6.2 Example Row (Original Data)**

scss

複製

市區道路,50,單路部分,直路,路段,一般車道(未劃分快慢車道),柏油,乾燥,無缺陷,無障礙物,無遮蔽物,良好,無號誌,無號誌,行車分向線,附標記,未繪設車道線,未繪設快慢車道分隔線,有,車與車,其他,駕駛者,違反其他標誌(線)禁制,死亡0;受傷1,1,小客車(含客、貨兩用),自用,男,46,繫安全帶(使用幼童安全椅),未使用,車的狀態,其他,汽車,左側車身,,駕駛者,違反其他標誌(線)禁制,否,120.624083,24.209466

* In the original data, attributes like weather, road condition, signal type, etc. are given as text. These are mapped to integers in the processed dataset.

**7. Usage Notes and Limitations**

1. **Data Completeness**:
   * Some rows are filtered out if they do not fit the mapping criteria (e.g., missing or unexpected text in 天候名稱).
2. **Integer Codes**:
   * The numeric codes are custom mappings. Make sure to keep the mapping dictionary accessible for interpretation.
3. **Granularity**:
   * Each row is at the accident level, so aggregated analysis (e.g., monthly or yearly trends) requires grouping by date or other attributes.
4. **Privacy**:
   * Personal identifying information (e.g., names, license plates) is not included in the dataset. Only aggregated counts of people or vehicles are provided.

**8. Contact and Citation**

* **Citation**: If you use or publish findings from this dataset, please cite the Taiwan Government Open Data Portal and reference the dataset [政府資料開放平台, dataset ID: 12197](https://data.gov.tw/dataset/12197).
* **Contact**: For questions about the original data structure, contact the data provider listed on the government open data page. For questions regarding the processing steps (mapping to integers, data cleaning), refer to this documentation or the accompanying code.

**Final Remarks**

This documentation should give you a clear understanding of the dataset’s structure, provenance, processing steps, and content. If you need further details—such as additional mappings, year-by-year record counts, or a breakdown of accident severities—feel free to expand the relevant sections accordingly.