# Applied Data Science Capstone project

## Problem definition

Hundreds of car accidents occur in major cities every day, some are minor accidents which result in minor property damage and some are major accidents resulting in of multiple fatalities.

There are a wide range of factors affecting the frequency and severity of accidents in any particular location: speed, vehicle type, vehicle condition, road condition, weather, human factors, lighting condition to name a few.

It will be very hard to predict the possibility and severity of accidents with high accuracy due to unpredictability of the affecting factors without using historical data and machine learning technics.

In this project we will use crash data for the last five years for Melbourne Australia provided by the Department of Transport of the state of Victoria to predict the locations that have a high chance of having serious crashes based on given conditions like day of the year and day of the week, hour, weather conditions, etc.

## Background

Victorian government has actively been planning to reduce the number of fatalities due to vehicle crashes in last 10 years, Australian government issues annual tables [[1]](#endnote-1) summarizing the number of road fatalities in each state.

The table indicates that there has been a 1.5% decrease in the number of fatalities in Victoria during the period of 2010-2019, and the number of fatalities in the state in 2019 is 270.

The annual fatality rate per 100,000 population at the same period has been reduced from 5.3 in 2010 to 4.1 in 2019.

With a focus on metropolitan Melbourne, increasing the effectiveness of the resources will have a huge effect on the number of serious crashes and the reduction of fatalities and injuries. As well as reducing the budget allocated to road safety.

## Data

Victorian “Department of Transport” issues information regarding crash data every year, and also a csv file containing the crash information in metropolitan Malborne in the last five years [[2]](#endnote-2). And it can be accessed [here](https://discover.data.vic.gov.au/dataset/crashes-last-five-years/resource/c08d93b2-6721-4f03-940a-65af88b844de)[[3]](#endnote-3).

The table includes about 195,000 records of the accidents in Metropolitan Melbourne in the last five years,

There are 63 attributes (Columns) for each record.

A brief summary of the attributes:

|  |  |
| --- | --- |
| FIELD NAME | FIELD DEFINITION |
| ABS\_CODE | Australian Bureau of Statistics classification of incidents |
| ACCIDENT\_DATE | Accident Date |
| ACCIDENT\_NO | Accident Number |
| ACCIDENT\_STATUS | Accident Status |
| ACCIDENT\_TIME | Accident Time |
| ACCIDENT\_TYPE | Accident Type |
| ALCOHOL\_RELATED | Alcohol Related Crashes BAC>0.001 and road user type=driver,rider,cyclist,pedestrian |
| ALCOHOLTIME | Incidents occurred within Road Crash Information System Definition of Alcohol Times |
| BICYCLIST | Number of pedal BICYCLISTs involved in the crash. |
| DAY\_OF\_WEEK | Day of week |
| DCA\_CODE | Definition for Classifying Accident. Link to DCA Chart and Sub DCA Codes https://vicroads-public.sharepoint.com/InformationAccess/Shared%20Documents/Road%20Safety/Crash/Accident/DCA\_Chart\_and\_Sub\_DCA\_Codes.PDF |
| DEG\_URBAN\_ALL | DEG\_URBAN\_ALL provides the type of urbanised area for the crash site in more detail. In some cases where a crash occurred on a border, several names will occur. |
| DEG\_URBAN\_NAME | DEG\_URBAN\_NAME provides the type of urbanised area for the crash site. |
| DIVIDED | DIVIDED is a character field that should indicate whether the crash occurred on a divided portion of road. |
| DIVIDED\_ALL | DIVIDED\_ALL is a character field that should indicate whether the crash occurred on a divided portion of road. |
| DRIVER | Number of DRIVERS involved in the crash. |
| FATALITY | Number of persons killed in the crash. |
| FEMALES | Total females involved in the crash. |
| HEAVYVEHICLE | Number of heavy vehicles involved in the crash. |
| HIT\_RUN\_FLAG | Indicates whether or not the crash was a hit-run accident. |
| INJ\_OR\_FATAL | Total Casualties - Total Persons Killed or Injured |
| Latitude | GDA94 Latitude coordinate - Decimal Degrees |
| LATITUDE | Geographical coordinates |
| LGA\_NAME | LGA\_NAME is a character field contains the LGA name. |
| LGA\_NAME\_ALL | LGA\_NAME\_ALL is a character field contains the name of the Local Government Area (LGA) in which the crash occurred. In some cases where a crash occurred on a border, several LGA names will occur. Unincorporated areas (usually Alpine resorts) will occur in brackets. |
| LIGHT\_CONDITION | Indicates the light condition or level of brightness at the time of the accident. |
| Longitude | GDA94 Longitude coordinate - Decimal Degrees |
| MALES | Total males involved in the crash. |
| MOTORCYCLE | Number of motorcycles involved in the crash. |
| MOTORIST | Number of MOTORCYCLISTS involved in the crash. |
| NO\_OF\_VEHICLES | Number of vehicles involved in the crash. |
| NODE\_ID | NODE\_ID is an integer field that uniquely identifies the accident node. It should start with 1 and be incremented by one when a new accident location is identified. |
| NODE\_TYPE | NODE\_TYPE is a character field indicates location type identified by the RCIS spatial system. |
| NONINJURED | Total persons involved but not injured in crash. |
| OLD\_DRIVER | Number of 65 years and older drivers involved in the crash. |
| OLD\_PEDESTRIAN | Number of 65 years and older pedestrians involved in the crash. |
| OTHERINJURY | Number of Persons injured but not classed as seriously injured in the crash. |
| PASSENGER | Number of Vehicle PASSENGERS involved in the crash. |
| PASSENGERVEHICLE | Number of passenger vehicles involved in the crash. |
| PED\_CYCLIST\_13\_18 | Number of 13 to 18 year old pedestrians and cyclists involved in the crash. |
| PED\_CYCLIST\_5\_12 | Number of 5 to 12 year old pedestrians and cyclists involved in the crash. |
| PEDESTRIAN | Number of Vehicle pedestrians involved in the crash. |
| PILLION | Number of Pillion Passengers involved in the crash. |
| POLICE\_ATTEND | POLICE\_ATTEND is a character field indicates whether the police attended the scene of the accident or not. |
| PUBLICVEHICLE | Number of Public Transport Vehicles (primarily trams and buses) involved in the crash. |
| REGION\_NAME | REGION\_NAME is a character field contains the VicRoads region name. |
| REGION\_NAME\_ALL | REGION\_NAME\_ALL is a character field contains the name of the VicRoads region in which the crash occurred. In some cases, where a crash occurred on a border, several VicRoads region names will occur. |
| RMA | Road Management Act (2004) classification |
| RMA\_ALL | Road Management Act road classification. With crashes occurring at junctions of road types providing combinations of types. |
| ROAD\_GEOMETRY | ROAD\_GEOMETRY is a character field indicates the layout of the road where the accident occurred. |
| RUN\_OFFROAD | Whether the crash involves a vehicle running off the road. |
| SERIOUSINJURY | Number of Persons seriously injured in the crash. Any person taken to hospital more likely to be classed as a serious injury. |
| SEVERITY | SEVERITY is a character field indicates VicRoads estimation of the severity or seriousness of the accident, based on the POLICE\_SEVERITY field. |
| SPEED\_ZONE | SPEED\_ZONE is a character field indicates the speed zone at the location of the accident. The speed zone is generally assigned to the main vehicle involved. |
| SRNS | Road on which the crash occurred classified by the State-wide Route Numbering Scheme (SRNS). 'M' roads provide a consistent high standard of driving conditions, with divided carriageways, four traffic lanes, sealed shoulders and line marking that is easily visible in all weather conditions. 'A' roads provide a similar high standard of driving conditions on a single carriageway. 'B' roads are sealed roads, wide enough for two traffic lines, with good centre line and edge line marking, shoulders, and a high standard of guidepost delineation. 'C' roads are generally two lane sealed roads with shoulders. Other roads are not classified. |
| SRNS\_ALL | Road on which the crash occurred classified by the State-wide Route Numbering Scheme (SRNS) with those occurring at junctions reflecting both types |
| STAT\_DIV\_NAME | STAT\_DIV\_NAME is a character field indicating the Metro Melbourne or Country region where the crash occurred. |
| TOTAL\_PERSONS | Total number of persons involved in the crash. |
| UNKNOWN | Number of persons involved in crash not classified into a known category of road user by police report. |
| UNLICENCSED | Unlicensed Drivers(road\_user\_type= driver & License Type =7 OR License Status< /> 9 and V Valid and Not Applicable) |
| VICGRID\_X | VICGRID\_X is a field indicating the grid reference in the x direction to provide a location reference for the crash using the VICGRID 1994 co-ordinate system. |
| VICGRID\_Y | VICGRID\_Y is a field indicating the grid reference in the y direction to provide a location reference for the crash using the VICGRID 1994 co-ordinate system. |
| YOUNG\_DRIVER | Number of 18-25 year old young drivers involved in the crash. |

Based on attribute description, 24 of the attributes were selected for this data analysis. These attributes seem to affect the severity of the crash.

|  |  |
| --- | --- |
| FIELD NAME | FIELD DEFINITION |
| ACCIDENT\_TIME | Accident Time |
| ACCIDENT\_TYPE | Accident Type |
| ALCOHOL\_RELATED | Alcohol Related Crashes BAC>0.001 and road user type=driver,rider,cyclist,pedestrian |
| ALCOHOLTIME | Incidents occurred within Road Crash Information System Definition of Alcohol Times |
| BICYCLIST | Number of pedal BICYCLISTs involved in the crash. |
| DAY\_OF\_WEEK | Day of week |
| HEAVYVEHICLE | Number of heavy vehicles involved in the crash. |
| Latitude | GDA94 Latitude coordinate - Decimal Degrees |
| LATITUDE | Geographical coordinates |
| LIGHT\_CONDITION | Indicates the light condition or level of brightness at the time of the accident. |
| Longitude | GDA94 Longitude coordinate - Decimal Degrees |
| MOTORCYCLE | Number of motorcycles involved in the crash. |
| NO\_OF\_VEHICLES | Number of vehicles involved in the crash. |
| NODE\_TYPE | NODE\_TYPE is a character field indicates location type identified by the RCIS spatial system. |
| OLD\_DRIVER | Number of 65 years and older drivers involved in the crash. |
| OLD\_PEDESTRIAN | Number of 65 years and older pedestrians involved in the crash. |
| PED\_CYCLIST\_13\_18 | Number of 13 to 18 year old pedestrians and cyclists involved in the crash. |
| PED\_CYCLIST\_5\_12 | Number of 5 to 12 year old pedestrians and cyclists involved in the crash. |
| PEDESTRIAN | Number of Vehicle pedestrians involved in the crash. |
| RMA | Road Management Act (2004) classification |
| ROAD\_GEOMETRY | ROAD\_GEOMETRY is a character field indicates the layout of the road where the accident occurred. |
| SEVERITY | SEVERITY is a character field indicates VicRoads estimation of the severity or seriousness of the accident, based on the POLICE\_SEVERITY field. |
| SPEED\_ZONE | SPEED\_ZONE is a character field indicates the speed zone at the location of the accident. The speed zone is generally assigned to the main vehicle involved. |
| STAT\_DIV\_NAME | STAT\_DIV\_NAME is a character field indicating the Metro Melbourne or Country region where the crash occurred. |
| YOUNG\_DRIVER | Number of 18-25 year old young drivers involved in the crash. |

The severity of an accident has been recorded in the “Severity” attribute and can be either of the following:

* Other injury accident
* Serious injury accident
* Non injury accident
* Fatal accident

The csv table will be converted to a Pandas DataFrame object and cleaned for further process. We will then normalize the data to prevent any biases and then will use several different trained machine learning methods to predict a future accident based on given data, the accuracy of these methods will be compared and the best method will be used.

1. <https://www.bitre.gov.au/sites/default/files/documents/annual_2019_tablesonly.xlsx> [↑](#endnote-ref-1)
2. <https://discover.data.vic.gov.au/dataset/crashes-last-five-years> [↑](#endnote-ref-2)
3. <https://discover.data.vic.gov.au/dataset/crashes-last-five-years/resource/c08d93b2-6721-4f03-940a-65af88b844de> [↑](#endnote-ref-3)