Predicting the severity of accidents in metropolitan Melbourne

Mohsen Rasekhi

October 12, 2020

1 Introduction

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] 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.

Interest

The prediction model will help first responders to the accident to be prepared with right amount or resources and give them the ability to prioritize them based on the information they receive about the accident.

2 DATA ACQUISITION AND CLEANING

Data source

Victorian “Department of Transport” issues information regarding crash data in the entire state of Victoria every year, and also a csv file containing the crash information in the last five years it can be accessed here[2].

The table includes about 75,000 records of the accidents in Victoria in the last five years,

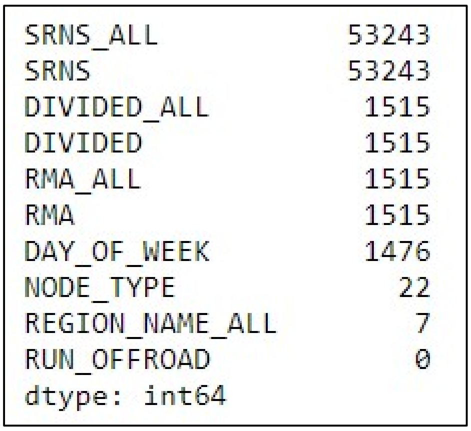
The data is comprehensive and it includes 63 attributes for each accident a summary of data description is provided at the end of this report.

Data cleaning

The data consists of 63 attributes with some missing data, there are a number of attributes that are not relevant and will need to be deleted these attributes include the ones that have too many unique values or the values that their presence defines the severity of the accident and is directly correlated to it, like if the accident involved any fatalities or injured people.

Missing data

the following attributes have missing values in them



```
SRNS_ALL            53243
SRNS                53243
DIVIDED_ALL          1515
DIVIDED              1515
RMA_ALL              1515
RMA                  1515
DAY_OF_WEEK          1476
NODE_TYPE              22
REGION_NAME_ALL         7
RUN_OFFROAD             0
dtype: int64
```

Attributes SRNS_ALL and SRNS were deleted due to high number of missing values.

For the rest of the attributes, the records with missing data will be deleted due to small number of missing records compared to the total number of records

Unique values

I have decided to remove any attributes with more than 105 unique values (this number was chosen to prevent ‘LGA_NAME’ attribute from being deleted.

I also deleted the following attributes because they were correlated with the existing data or were irrelevant:

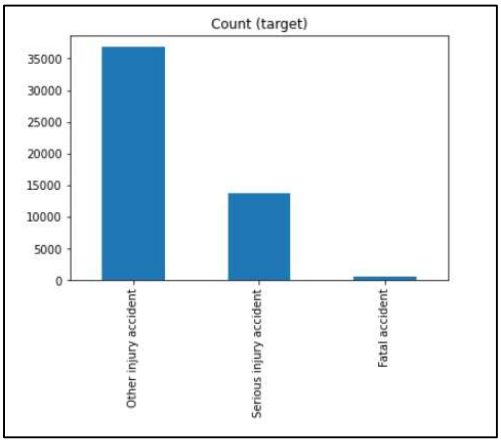
Deleted attributes

['OTHERINJURY','SERIOUSINJURY','UNKNOWN','NONINJURED','FATALITY',"INJ_OR_FATAL",'POLICE_ATTEND','ACCIDENT_STATUS','ABS_CODE','REGION_NAME',"DEG_URBAN_NAME",'DEG_URBAN_ALL','REGION_NAME_ALL','SRNS','RMA','DIVIDED','STAT_DIV_NAME','LGA_NAME_ALL','DCA_CODE','MALES','FEMALES','TOTAL_PERSONS','PASSENGER','ACCIDENT_TYPE']

Data resemblance

The SEVERITY had 3 unique values;

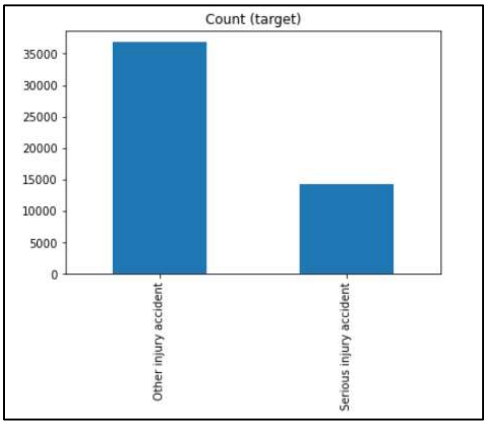
Other injury accident	36817	records
Serious injury accident	13701	records
Fatal accident	608	records



The number of records for each Severity category indicates that there is a large imbalance between them, and in order to prevent bias in the prediction results, this imbalance will need to be resolved.

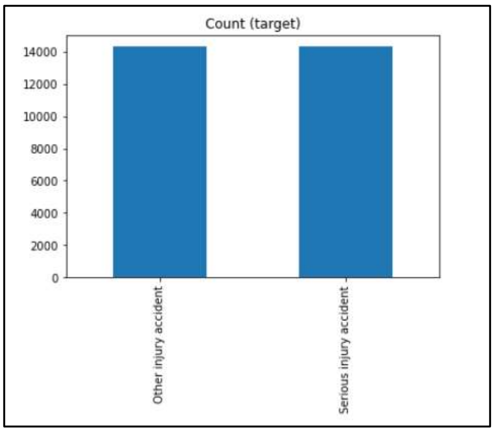
First, we combine the “Fatal accident’ and ‘Serious injury accident’ to reduce the number of categories to two.

Other injury accident	36817	records
Serious injury accident	14309	records



And then we use a resemblance technique called Under-Sampling to randomly choose the same number of records as the attribute with less records (in this case we randomly choose 14838 records from the “Other injury accident” and ignore the rest of the data.

Now the database has 2x14309= 28618 records instead of original 75000 records.



Reducing accident time categories

The time of accident has too many different values, to reduce them, we will extract the “Hour” of the accident so that there are 24 different values in this attribute.

Next step is to change all the categorical values to numerical values so that the models can analyze the data. This will be done by using LabelEncoder method in Sklearn.

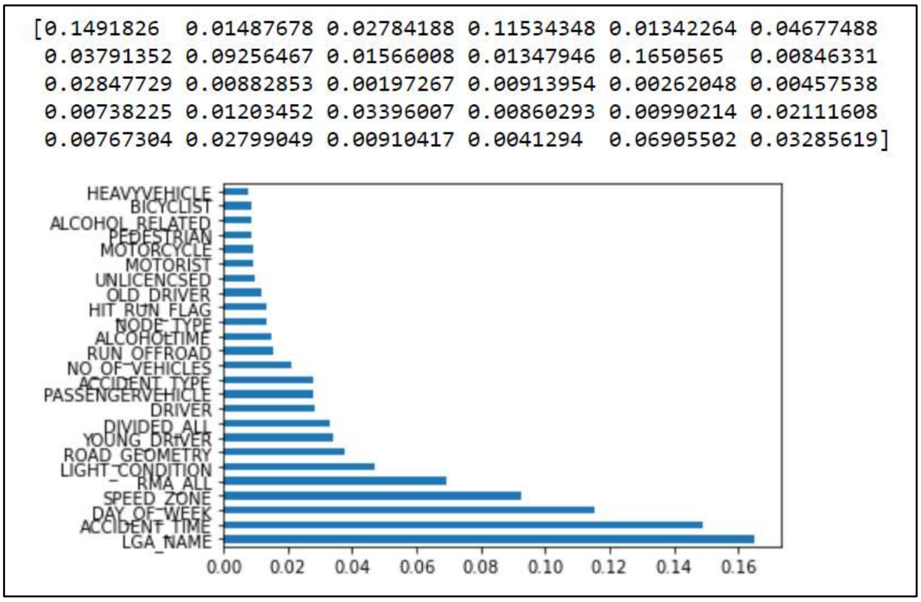
Normalization

To give the different attributes a similar weight in the model, we will use StandarScaler method in Sklearn.

3 FEATURE SELECTION

After removing all the unwanted attributes, there are still 31 attributes left, to be able to reduce them even further we will use Sklearn feature selection method ExtraTreesClassifier, this method analyses the data and ranks the features that are more relevant to the target.

We will select 25 attributes out of existing 29, the features as derived by the ExtraTreesClassifier method are shown below:



```
[0.1491826  0.01487678 0.02784188 0.11534348 0.01342264 0.04677488
 0.03791352 0.09256467 0.01566008 0.01347946 0.1650565  0.00846331
 0.02847729 0.00882853 0.00197267 0.00913954 0.00262048 0.00457538
 0.00738225 0.01203452 0.03396007 0.00860293 0.00990214 0.02111608
 0.00767304 0.02799049 0.00910417 0.0041294  0.06905502 0.03285619]
```

4 Exploratory Data Analysis

Target variable

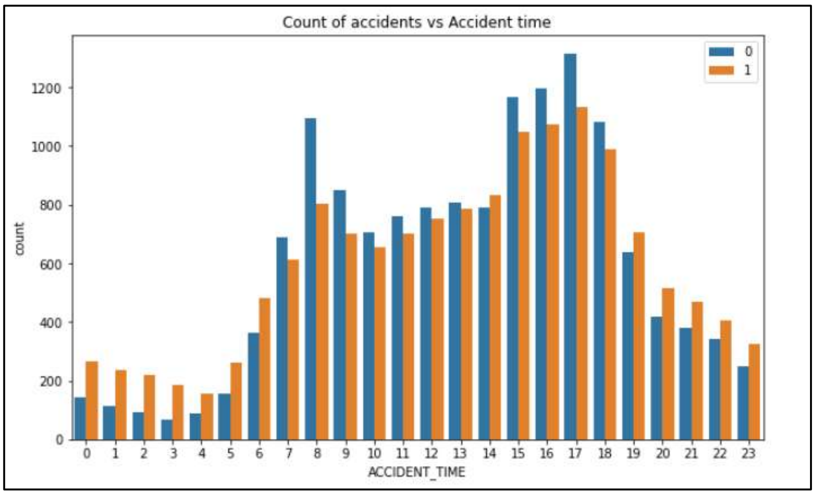
The target variable is the "SEVERITY" of the accident which originally had 3 different values and was reduced to two variables to deal with the imbalance in the number of records. The severity of the accident depends on various factors including if there were any fatalities or injured people in the accident or the degree of damage to the property.

Relationships,

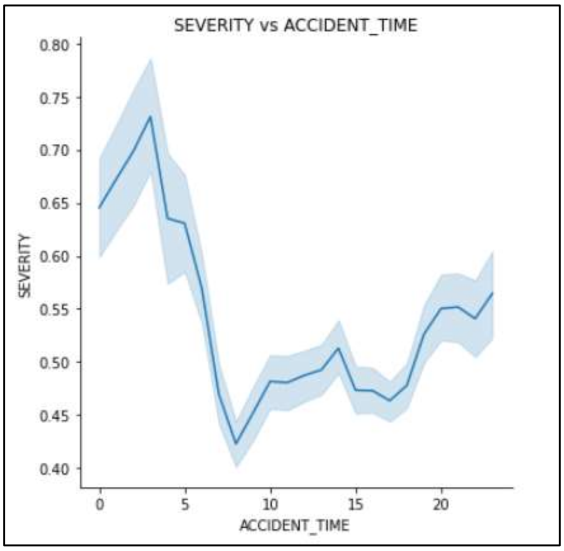
The relationship of some attributes vs severity of accident are shown below

Relationship between ACCIDENT_TIME and SEVERITY

Although in general the number of accidents increase during the day as shown below,

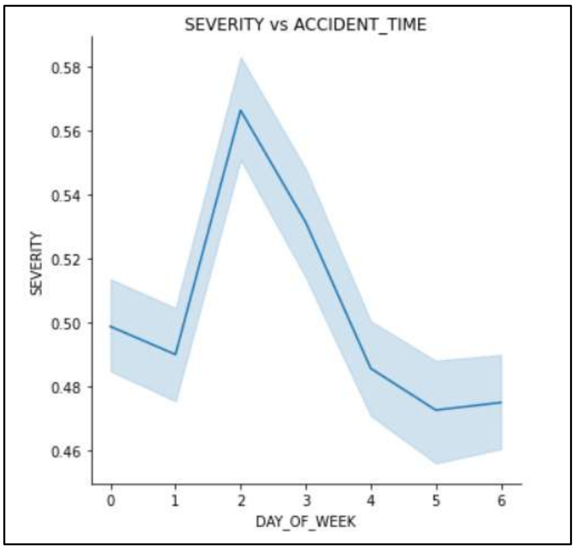


The severity of accidents follows a different pattern, if an accident happens in the early hours of the morning in is more likely to be a serious accident than when in happens during other times



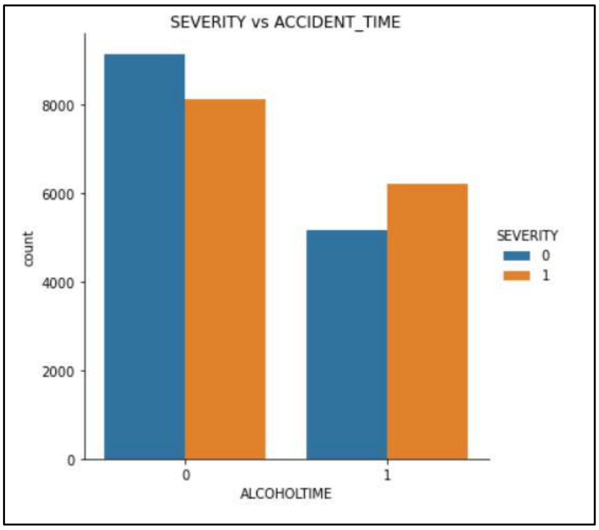
Relationship between DAY_OF WEEK and SEVERITY

The data shows that day of week has an effect on the number of severe accidents



Relationship between Alcohol time and severity

The graph indicates that the proportion of serious accidents increases at alcohol time.



5 PREDICTIVE MODELING

The prediction that we are trying to achieve is a classification problem because we want to determine if an accident is serious or not serious,

For modelling I have used three different classification models and compared the accuracy factors.

The classification models used are:

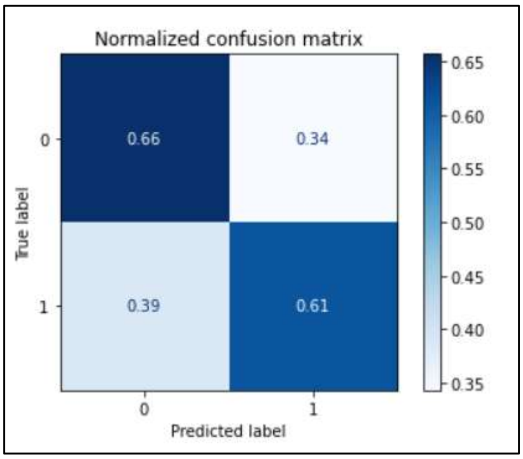
- Logistic regression Classifier
- KNeighbors Classifier
- RandomForest Classifier

A summary of accuracy result is shown in this table:

SUMMARY OF MODELLING ACCURACY RESULTS		
Classifier	Accuracy	
	Training set	Test set
Logistic regression	63%	63%
KNeighbors	63%	63%
RandomForest	99%	61%

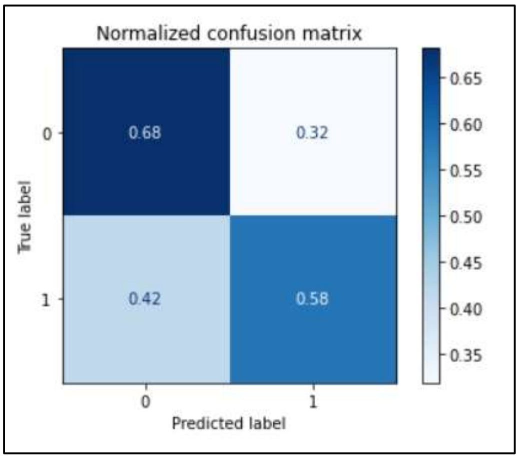
To further assess the accuracy and reliability of the results I used confusion matrix to see if the results are biased towards any of the severity categories.

Confusion matrix for Logistic regression:



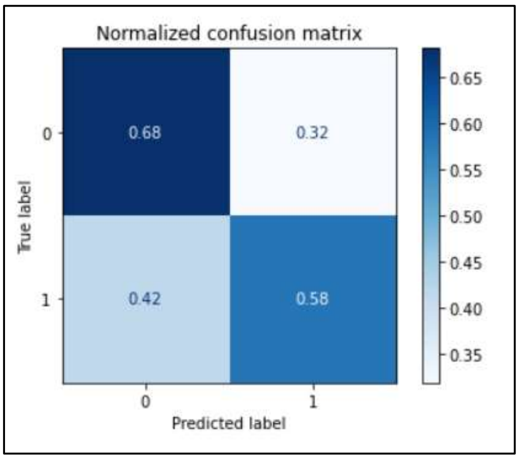
Confusion matrix for Logistic regression:

Confusion matrix for KNeighbors Classifier



Confusion matrix for KNeighbors Classifier

Confusion matrix for RandomForest Classifier



Confusion matrix for RandomForest Classifier

A summary of confusion matrix result is shown in this table:

SUMMARY OF CONFUSION MATRIX RESULTS		
Classifier	Accuracy	
	SEVERITY=0	SEVERITY=1
Logistic regression	66%	61%
KNeighbors	68%	58%
RandomForest	68%	58%

The results indicate that the Logistic regression model was more accurate in predicting “Severe Injury Accidents” and also it was much faster. Therefore, Logistic regression modelling was chosen.

6 CONCLUSION.

The accident severity is a highly unpredictable issue and a lot of unpredictable factors can influence the severity of an accident.

I realized that time of accident, alcohol and speed zones are some of the important factors as well as if the accident involved pedestrian and bicycles.

The model can be used to determine in a given condition, how likely an accident can be a serious one and first responders can be better prepared for it.

7 FUTURE DIRECTIONS

I was able to achieve 61% accuracy in determining the severity of an accident in my classification model.

Considering the highly unpredictability of severity of a car accident, I think that more relevant data that can influence the accident needs to be collected, including weather condition, model of the car, type of the vehicle, temperature, etc.

Another feature that can improve the accuracy would be creating a geographic grid zone with specific numbers allocate to each accident, this can be more useful than just the coordinates of the accident in determining if the location contributes to the severity of an accident.

Appendix 1: Attribute definition

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.

[1] https://www.bitre.gov.au/sites/default/files/documents/annual_2019_tablesonly.xlsx

[2] https://discover.data.vic.gov.au/dataset/crashes-last-five-years/resource/c08d93b2-6721-4f03-940a-65af88b844de