TDS3301 DATA MINING

Trimester 1, 2023/2024

PROJECT (30%)

Predictive modelling of accident severity and number of casualties in the UK

Project Member:

Student Name	Student ID
Chia Yu Zhang	1201101003
Yap Zhi Toung	1201100514
Gerald Godwin Lee Yong Lin	1201100324
Teo Hazel	1201100924

Contribution Table

Name	ID	Data Preprocessing	Feature Selection	Classification
Chia Yu Zhang	1201101003		✅	✅
Yap Zhi Toung	1201100514	✅		
Gerald Godwin Lee Yong Lin	1201100324		✅	
Teo Hazel	1201100924			✅

Abstract

This research project aimed to predict accident severity and the number of casualties in the UK using machine learning techniques. Road accidents pose significant human and economic risks globally, making them a critical concern. By exploring and identifying predictors of accident severity and casualty numbers in the UK through feature selection and classification models, the objective was to develop a predictive model that could assess the quality of their relationship.

This study utilised a comprehensive dataset on accident severity and casualty numbers in the United Kingdom. Descriptive analysis of the dataset illustrated the distribution among accident severity and casualty numbers in the UK. Classification models were built using Random Forest, Logistic Regression, K-Nearest Neighbors, and Decision Tree algorithms to make predictions on different target variables. The findings included evaluation metrics such as accuracy, precision, recall, and F1-score.

Keywords: Data Mining, Classification, Random Forest, Logistic Regression, K-Nearest Neighbour and Decision Tree.

Table of Content

1. Introduction

Road accidents represented a severe and widespread problem that had a significant impact on public safety and economic stability both internationally and in the United Kingdom. Accident severity and the number of casualties had far-reaching effects, including fatalities and significant financial costs, and demanded attention. Understanding the impact of road-related factors, environment-related factors, driver-related factors, and accident-related factors' relationships and predicting which factor caused larger amounts of accident severity and the number of casualties were crucial objectives.

In this research project, a large-scale dataset primarily capturing road accidents in the UK between 1979 and 2015 was provided for us to gather important information for predictive modelling on accident severity and the number of casualties. The classification models were performed to investigate, explore data, and predict accident severity and the number of casualties, visualising patterns and demonstrating causal links. In conclusion, different machine learning models such as random forest, logistic regression, and decision tree were implemented to evaluate and test the accuracy, precision, recall, F1-score, confusion matrix, and classification report. Comparison was made based on the evaluation results to find the best model.

2. Related Works

Antonio Comi, Antonio Polimeni, and Chiara Balsamo (2021) identified which data mining techniques were most apt for examining road accidents and successfully categorized the significant causes and recurring patterns of these incidents through a descriptive approach for accident data of the 15 districts of Rome Municipality collected from 2016 to 2019. The authors aimed to discover the most recurrent patterns of road accidents by means of descriptive analysis. K-Means algorithm and Kohonen network were evaluated in this experiment for descriptive analysis, while decision trees and neural networks were evaluated for predictive analysis. In conclusion, hybrid prediction approaches should be investigated so that they are effective for both statistical and machine learning models in accident prediction. For these methods, they can be applied to the project to achieve the target by using the related factors. The patterns can be illustrated, and some interesting observations can be found. Moreover, efficient accuracy, MSE, and RMSE results can be achieved.

Eboli et al., 2020 proposed a paper that investigated the factors influencing accident severity. The authors concluded that the driver's characteristics, road and vehicle features, accident characteristics, and weather factors affected the accident severity. To analyse the main factor causing accident severity, they used data on road accidents that occurred in Italy during 2016 and divided it into four categories: road, external environment, driver, and accident. They used binary logistic regression to evaluate and test the relationship of the impact of road-related factors, environment-related factors, driver-related factors, and accident-related factors. In the project, the target is to analyse which factor in the datasets will affect accident severity and the number of casualties, but the dataset is too large for the team to perform the RFE process. Therefore, in this report, the RFE process can be streamlined by selecting attributes from the location, weather, driver behaviour, vehicle status, and time categories.

The proposed paper by (Priya et al.,2018) used advanced data mining and predictive modelling methods to help learn more about traffic safety and crash analysis. To determine complicated crash patterns, the study employed strong algorithms like the Apriori algorithm, K-modes clustering, and decision tree classification, all tailored to the UK's geography. It discussed choosing the right features, with a focus on finding the most important variables for

accurate prediction tasks. A temporal study showed how accidents changed over time. The research effort carefully looked at machine learning techniques and their pros and cons, giving a more complete picture of how useful they were for predicting accident outcomes. Looking at UK government records and policies on road safety provided a regulatory view and showed how actions taken by the government might affect how severe accidents were and how many people died. It dealt with problems that came up when collecting and analysing data, and it changed as methods changed to keep an open mind. Overall, (Priya et al.,2018) paper fits in with other research that had already been done and gives a complete and in-depth look at predictive modelling in traffic safety and crash analysis.

The paper titled “Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach” by Yassin, S.S., & Pooja (2020) and this current research project shared common objectives regarding the predictive modelling of accident severity, albeit within different geographic contexts and methodological frameworks. While the paper focused on road accidents in Addis Ababa, Ethiopia, this project examined data spanning several decades from road accidents in the United Kingdom. Both endeavours utilised machine learning algorithms to analyse extensive datasets and identify factors influencing accident severity. Notably, the paper employed a hybrid approach involving K-means and Random Forest algorithms, whereas this project explored various models including Random Forest, Logistic Regression, K-Nearest Neighbors, and Decision Trees. Despite these methodological differences, both studies emphasised the importance of understanding the complex interplay between factors such as driver experience, environmental conditions, and vehicle characteristics in determining accident severity. Furthermore, both initiatives sought to inform road safety strategies and interventions by providing actionable insights derived from predictive modelling. Thus, while the paper and this project diverged in methodology and dataset characteristics, they contributed valuable insights to the broader discourse on road safety and accident prevention within their respective contexts.

3. About the Dataset

The road accident incident dataset captured road accidents in the UK between 1979 and 2015, containing 70 attributes and approximately 250k rows, as shown in Table 2.1. The dataset was provided by Akshay Babbar in 2017 and was available for download on Kaggle, an open-source website, in CSV format (Road Accidents Incidence (kaggle.com), 10/1/2024). The author also provided an Excel file named Road-Accident-Safety-Data-Guide.xls to other researchers to understand the attributes recorded inside the dataset, as shown in Table 2.1 data columns.

Table 2.1 Road Accident Incident Datasets.

Attributes Name	Description	Data
Accident index	A unique identifier for each road accident incident.	Specific number = accident id
vehicle_reference	A reference number or identifier for the vehicle involved in the accident	The identifier of the vehicle represented.
vehicle_type	Type of vehicle involved in an accident.	1=Pedal cycle 2=Motorcycle 50cc and under 3=Motorcycle 125cc and under 4=Motorcycle over 125cc and up to 500cc 5=Motorcycle over 500cc 8=Taxi/Private hire car 9=Car 10=Minibus (8 - 16 passenger seats) 11=Bus or coach (17 or more pass seats) 16=Ridden horse 17=Agricultural vehicle

		18=Tram 19=Van / Goods 3.5 tonnes mgw or under 20=Goods over 3.5t. and under 7.5t 21=Goods 7.5 tonnes mgw and over 22=Mobility scooter 23=Electric motorcycle 90=Other vehicle 97=Motorcycle - unknown cc 98=Goods vehicle - unknown weight -1=Data missing or out of range
towing_and_articulation	Indicates whether the vehicle was towing or articulated something.	0=No tow/articulation 1=Articulated vehicle 2=Double or multiple trailer 3=Caravan 4=Single trailer 5=Other tow -1=Data missing or out of range
vehicle_manoeuvre	Describes the manoeuvre or action the vehicle was performing at the time of the accident	1=Reversing 2=Parked 3=Waiting to go - held up 4=Slowing or stopping 5=Moving off 6=U-turn 7=Turning left 8=Waiting to turn left 9=Turning right

		10=Waiting to turn right 11=Changing lane to left 12=Changing lane to right 13=Overtaking moving vehicle - offside 14=Overtaking static vehicle - offside 15=Overtaking - nearside 16=Going ahead left-hand bend 17=Going ahead right-hand bend 18=Going ahead other -1=Data missing or out of range
vehicle_location-restricted_lane	Indicates if the vehicle was in a restricted area when the accident happened.	0=On main c'way - not in restricted lane 1=Tram/Light rail track 2=Bus lane 3=Busway (including guided busway) 4=Cycle lane (on main carriageway) 5=Cycleway or shared use footway (not part of main carriageway) 6=On lay-by or hard shoulder 7=Entering lay-by or hard shoulder 8=Leaving lay-by or hard shoulder 9=Footway (pavement) 10=Not on carriageway

		-1=Data missing or out of range
junction_location	Specifies the location of the accident related to the junction.	0=Not at or within 20 metres of junction 1=Approaching junction or waiting/parked at junction approach 2=Cleared junction or waiting/parked at junction exit 3=Leaving roundabout 4=Entering roundabout 5=Leaving main road 6=Entering main road 7=Entering from slip road 8=Mid Junction - on roundabout or on main road -1=Data missing or out of range
skidding_and_overturning	Indicates whether skidding or overturning occurred during the accident	0=None 1=Skidded 2=Skidded and overturned 3=Jackknifed 4=Jackknifed and overturned 5=Overturned -1=Data missing or out of range
hit_object_in+carriageway	Indicates whether the vehicle collided with an object within the carriageway (roadway).	0= None 1= Previous accident 2= Road works 4= Parked vehicle 5= Bridge (roof)

		6= Bridge (side) 7= Bollard or refuge 8= Open door of vehicle 9= Central island of roundabout 10= Kerb 11=Other object 12= Any animal (except ridden horse) -1= Data missing or out of range
vehicle_leaving_carriageway	Indicates whether the vehicle involved in the accident left the carriageway.	0=Did not leave carriageway 1=Nearside 2=Nearside and rebounded 3=Straight ahead at junction 4=Offside on to central reservation 5=Offside on to centrl res + rebounded 6=Offside - crossed central reservation 7=Offside 8=Offside and rebounded -1=Data missing or out of range
hit_object_off_carriageway	Specifies whether the vehicle collided with an object outside the carriageway (roadway)	0=None 1=Road sign or traffic signal 2=Lamp post 3=Telegraph or electricity pole 4=Tree

		5=Bus stop or bus shelter 6=Central crash barrier 7=Near/Offside crash barrier 8=Submerged in water 9=Entered ditch 10=Other permanent object 11=Wall or fence -1=Data missing or out of range
1st_point_of_impact	Specifies the location of the first point of the impact on the vehicle	0=Did not impact 1=Front 2=Back 3=Offside 4=Nearside -1=Data missing or out of range
was_vehicle_left_hand_drive?	Indicates if the vehicle was left-hand drive (as opposed to right-hand drive).	1= No 2= Yes -1= Data missing or out of range
journey_purpose_of_driver	The reason for the driver's journey at the time of the accident.	1= Journey as part of work 2= Commuting to/from work 3= Taking pupil to/from school 4= Pupil riding to/from school 5= Other 6= Not known 15= Other/Not known (2005-10) -1=Data missing or out of range

sex_of_driver	The gender of the driver involved in the accident.	1=Male 2=Female 3=Not known -1=Data missing or out of range
age_of_driver	The age of the driver involved in the accident.	Number based on driver's age -1= Data missing or out of range
age_band_of_driver	Age band or category to which the driver belongs.	1 = 0-5 2 = 6 - 10 3 = 11 - 15 4 = 16 - 20 5 = 21 - 25 6 = 26 - 35 7 = 36 - 45 8 = 46 - 55 9 = 56 - 65 10 = 66 - 75 11 = Over 75 -1 = Data missing or out of range
engine_capacity_(cc)=	The engine capacity of the vehicle in cubic centimetres (cc).	Number based on engine capacity (cc) -1= Data missing or out of range
propulsion_code	Code indicating the type of propulsion used by the vehicle.	1 = Petrol 2 = Heavy oil 3 = Electric 4 = Steam 5 = Gas 6 =Petrol/Gas (LPG)

		7 = Gas/Bi-fuel 8 = Hybrid electric 9 = Gas Diesel 10 = New fuel technology 11 = Fuel cells 12 = Electric diesel M = Undefined
age_of_vehicle	The age of the vehicle involved in the accident.	Number based on vehicle age -1= Data missing or out of range
driver_imd_decile	The Index of Multiple Deprivation (IMD) decile of the driver's residence area.	1 = Most deprived 10% 2 = More deprived 10-20% 3 = More deprived 20-30% 4 = More deprived 30-40% 5 = More deprived 40-50% 6 = Less deprived 40-50% 7 = Less deprived 30-40% 8 = Less deprived 20-30% 9 = Less deprived 10-20% 10 = Least deprived 10% -1 = Data missing or out of range
driver_home_area_type	The type of area (urban or rural) where the driver resides.	1 = Urban area 2 = Small town 3 = Rural -1 = Data missing or out of range
vehicle_imd_decile	The IMD decile for the area where the vehicle is registered.	1 = Most deprived 10% 2 = More deprived 10-20% 3 = More deprived 20-30% 4 = More deprived 30-40% 5 = More deprived 40-50%

		6 = Less deprived 40-50% 7 = Less deprived 30-40% 8 = Less deprived 20-30% 9 = Less deprived 10-20% 10 = Least deprived 10% -1 = Data missing or out of range
NUmber_of_Casualities_unique_to_accident_index	The number of casualties unique to the accident index.	Number based on casualties count
No_of_Vehicles_involved_unique_to_accident_index	The number of vehicles involved is unique to the accident index.	Number based on vehicle count
location_easting_osgr	Easting coordinate of the accident location in Ordnance Survey Grid Reference.	Easting coordinate
location_northing_osgr	Northing coordinate of the accident location in Ordnance Survey Grid Reference.	Northing coordinate
longitude	The geographical longitude of the accident location.	Longitude number
latitude	The geographical latitude of the accident location.	Latitude number
police_force	The police force that attended the accident scene.	1= Metropolitan Police 3= Cumbria 4= Lancashire 5= Merseyside 6= Greater Manchester 7= Cheshire 10= Northumbria 11= Durham 12=North Yorkshire 13=West Yorkshire 14=South Yorkshire 16=Humberside

		17=Cleveland 20=West Midlands 21=Staffordshire 22=West Mercia 23=Warwickshire 30=Derbyshire 31=Nottinghamshire 32=Lincolnshire 33=Leicestershire 34=Northamptonshire 35=Cambridgeshire 36=Norfolk 37=Suffolk 40=Bedfordshire 41=Hertfordshire 42=Essex 43=Thames Valley 44=Hampshire 45=Surrey 46=Kent 47=Sussex 48=City of London 50=Devon and Cornwall 52=Avon and Somerset 53=Gloucestershire 54=Wiltshire 55=Dorset 60=North Wales 61=Gwent 62=South Wales 63=Dyfed-Powys 91=Northern 92=Grampian 93=Tayside 94=Fife 95=Lothian and Borders

		96=Central 97=Strathclyde 98=Dumfries and Galloway
accident_severity	The severity of the accident (fatal, serious, or slight).	1=Fatal 2=Serious 3=Slight
number_of_vehicles	The total number of vehicles involved in the accident.	Number based on vehicle count
number_of_casualties	The total number of casualties in the accident.	Number based on casualties count
date	The date when the accident occurred.	(DD/MM/YYYY)
day_of_week	The day of the week when the accident occurred.	1=Monday 2=Tuesday 3=Wednesday 4=Thursday 5=Friday 6=Saturday 7=Sunday
time	The time of the day when the accident occurred.	(HH:MM)
local_authority_(district)	The local authority district where the accident occurred.	400++ variables
local_authority_(highway)	The local authority responsible for the highway where the accident occurred.	200++ variables
1st_road_class	The classification of the first road involved in the accident.	1=Motorway 2=A(M) 3=A 4=B 5=C

		6=Unclassified
1st_road_number	The number of the first road involved in the accident.	0 = road number is not applicable or not recorded. Specific number = correspond to specific roads or locations where accidents have been recorded.
road_type	The type of road where the accident occurred.	1=Roundabout 2=One way street 3=Dual carriageway 6=Single carriageway 7=Slip road 9=Unknown 12=One way street/Slip road -1=Data missing or out of range
speed_limit	The speed limit on the road where the accident occurred.	Number based on speed limit (km/h)
junction_detail	Details about the type of junction where the accident occurred.	0=Not at junction or within 20 metres 1=Roundabout 2=Mini-roundabout 3=T or staggered junction 5=Slip road 6=Crossroads 7=More than 4 arms (not roundabout) 8=Private drive or entrance 9=Other junction -1=Data missing or out of range

junction_control	Control measures at the junction where the accident occurred.	0=Not at junction or within 20 metres 1=Authorised person 2=Auto traffic signal 3=Stop sign 4=Give way or uncontrolled -1=Data missing or out of range
2nd_road_class	The classification of the second road involved in the accident.	0=Not at junction or within 20 metres 1=Motorway 2=A(M) 3=A 4=B 5=C 6=Unclassified
2nd_road_number	The number of the second road involved in the accident.	0 = road number is not applicable or not recorded. Specific number = correspond to specific roads or locations where accidents have been recorded.
pedestrian_crossing-human_control	Presence of human-controlled pedestrian crossing at the accident site.	0=None within 50 metres 1=Control by school crossing patrol 2=Control by other authorised person -1=Data missing or out of range
pedestrian_crossing-physical_facilities	Physical facilities available at pedestrian crossings at the accident site.	0=No physical crossing facilities within 50 metres

		1=Zebra 4=Pelican, puffin, toucan or similar non-junction pedestrian light crossing 5=Pedestrian phase at traffic signal junction 7=Footbridge or subway 8=Central refuge -1=Data missing or out of range
light_conditions	Lighting conditions at the time of the accident.	1=Daylight 4=Darkness - lights lit 5=Darkness - lights unlit 6=Darkness - no lighting 7=Darkness - lighting unknown -1=Data missing or out of range
weather_conditions	Weather conditions at the time of the accident.	1=Fine no high winds 2=Raining no high winds 3=Snowing no high winds 4=Fine + high winds 5=Raining + high winds 6=Snowing + high winds 7=Fog or mist 8=Other 9=Unknown -1=Data missing or out of range
road_surface_conditions	Surface conditions of the road at the time of the accident.	1=Dry 2=Wet or damp 3=Snow 4=Frost or ice 5=Flood over 3cm. deep

		6=Oil or diesel 7=Mud -1=Data missing or out of range
special_conditions_at_site	Special conditions present at the accident site.	0=None 1=Auto traffic signal - out 2=Auto signal part defective 3=Road sign or marking defective or obscured 4=Roadworks 5=Road surface defective 6=Oil or diesel 7=Mud -1=Data missing or out of range
carriageway_hazards	Hazards present on the carriageway at the time of the accident.	0=None 1=Vehicle load on road 2=Other object on road 3=Previous accident 4=Dog on road 5=Other animal on road 6=Pedestrian in carriageway - not injured 7=Any animal in carriageway (except ridden horse) -1=Data missing or out of range
urban_or_rural_area	Whether the accident occurred in an urban or rural area.	1=Urban 2=Rural 3=Unallocated

did_police_officer_attend_scene_of_accident	Indicates whether a police officer attended the scene of the accident.	1=Yes 2=No 3=No - accident was reported using a self completion form (self rep only)
lsoa_of_accident_location	Lower Layer Super Output Area (LSOA) of the accident location.	LSOA codes assigned to each Lower Layer Super Output Area
casualty_reference	Reference number for the casualty involved in the accident.	The identifier of casualty represented.
casualty_class	Describe the classification of the casualty involved in the accident.	1=Driver or rider 2=Passenger 3=Pedestrian
sex_of_casualty	Specifies the gender of the casualty Involved in the accident	1=Male 2=Female -1=Data missing or out of range
age_of_casualty	Indicates the age of the casualty involved in the accident	0-120= 0-120 -1=Data missing or out of range
age_band_of_casualty	Groups casualties into age bands for analysis	1 =0 - 5 2 =6 - 10 3 =11 - 15 4 =16 - 20 5 =21 - 25 6 =26 - 35 7 =36 - 45 8 =46 - 55 9 =56 - 65 10 =66 - 75 11 =Over 75

		-1 =Data missing or out of range
casualty_severity	Indicates the severity of the casualty's injuries or condition as a result of the accident	1=Fatal 2=Serious 3=Slight
pedestrian_location	Describe the location of the pedestrian at the time of the accident	0=Not a Pedestrian 1=Crossing on pedestrian crossing facility 2=Crossing in zig-zag approach lines 3=Crossing in zig-zag exit lines 4=Crossing elsewhere within 50m. of pedestrian crossing 5=In carriageway, crossing elsewhere 6=On footway or verge 7=On refuge, central island or central reservation 8=In centre of carriageway - not on refuge, island or central reservation 9=In carriageway, not crossing 10=Unknown or other -1=Data missing or out of range
pedestrian_movement	Specifies the movement or action of the pedestrian at the time of the accident	0=Not a Pedestrian 1=Crossing from driver's nearside 2=Crossing from nearside - masked by parked or stationary vehicle

		3=Crossing from driver's offside 4=Crossing from offside - masked by parked or stationary vehicle 5=In carriageway, stationary - not crossing (standing or playing) 6=In carriageway, stationary - not crossing (standing or playing) - masked by parked or stationary ve... 7=Walking along in carriageway, facing traffic 8=Walking along in carriageway, back to traffic 9=Unknown or other -1=Data missing or out of range
car_passenger	Indicates whether the casualty was a passenger on a bus or coach involved in the accident.	0=Not car passenger 1=Front seat passenger 2=Rear seat passenger -1=Data missing or out of range
bus_or_coach_passenger	Indicates whether the casualty was a passenger in a bus or coach involved in the accident.	0=Not a bus or coach passenger 1=Boarding 2=Alighting 3=Standing passenger 4=Seated passenger -1=Data missing or out of range

pedestrian_road_maintenance_worker	Indicates whether the casualty was a pedestrian road maintenance worker.	0=No / Not applicable 1=Yes 2=Not Known -1=Data missing or out of range
casualty_type	Describes the type of casualty involved in the accident	0=Pedestrian 1=Cyclist 2=Motorcycle 50cc and under rider or passenger 3=Motorcycle 125cc and under rider or passenger 4=Motorcycle over 125cc and up to 500cc rider or passenger 5=Motorcycle over 500cc rider or passenger 8=Taxi/Private hire car occupant 9=Car occupant 10=Minibus (8 - 16 passenger seats) occupant 11=Bus or coach occupant (17 or more pass seats) 16=Horse rider 17=Agricultural vehicle occupant 18=Tram occupant 19=Van / Goods vehicle (3.5 tonnes mgw or under) occupant 20=Goods vehicle (over 3.5t. and under 7.5t.) occupant

		21=Goods vehicle (7.5 tonnes mgw and over) occupant 22=Mobility scooter rider 23=Electric motorcycle rider or passenger 90=Other vehicle occupant 97=Motorcycle - unknown cc rider or passenger 98=Goods vehicle (unknown weight) occupant
casualty_home_area_type	The type of area (urban or rural) where the casualty resides.	1 = Urban area 2 = Small town 3 = Rural -1 = Data missing or out of range
casualty_imd_decile	The IMD decile of the area where the casualty resides.	1 = Most deprived 10% 2 = More deprived 10-20% 3 = More deprived 20-30% 4 = More deprived 30-40% 5 = More deprived 40-50%

4. Data Preprocessing

4.1 Overview of Data

As seen in Figure 4.1.1, the output of the df.info() function provided details about a data frame with the name df. It displayed that there were 70 columns and 285,331 rows in the data frame. A variety of data types were represented in these columns: object (string), int64 (integer), and float64 (floating-point number). The number of non-missing values was represented by the corresponding non-null count for each column. Furthermore, the data frame's memory utilisation was given. As a result, the output provided a brief summary of the data frame's structure as well as the data kinds linked to each of its columns.

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 285331 entries, 0 to 285330
Data columns (total 70 columns):
 #   Column                                          Non-Null Count   Dtype
---  ------                                          --------------   -----
 0   accident_index                                  285331 non-null  object
 1   vehicle_reference                               285331 non-null  int64
 2   vehicle_type                                    285331 non-null  int64
 3   towing_and_articulation                         285331 non-null  int64
 4   vehicle_manoeuvre                               285331 non-null  int64
 5   vehicle_location-restricted_lane                285331 non-null  int64
 6   junction_location                               285331 non-null  int64
 7   skidding_and_overturning                        285331 non-null  int64
 8   hit_object_in_carriageway                       285331 non-null  int64
 9   vehicle_leaving_carriageway                     285331 non-null  int64
 10  hit_object_off_carriageway                      285331 non-null  int64
 11  1st_point_of_impact                             285331 non-null  int64
 12  was_vehicle_left_hand_drive?                    285331 non-null  int64
 13  journey_purpose_of_driver                       285331 non-null  int64
 14  sex_of_driver                                   285331 non-null  int64
 15  age_of_driver                                   285331 non-null  int64
 16  age_band_of_driver                              285331 non-null  int64
 17  engine_capacity_(cc)                            285331 non-null  int64
 18  propulsion_code                                 285331 non-null  int64
 19  age_of_vehicle                                  285331 non-null  int64
 20  driver_imd_decile                               285331 non-null  int64
 21  driver_home_area_type                           285331 non-null  int64
 22  vehicle_imd_decile                              285331 non-null  int64
 23  NUmber_of_Casualities_unique_to_accident_index  285331 non-null  int64
 24  No_of_Vehicles_involved_unique_to_accident_index 285331 non-null  int64
 25  location_easting_osgr                           285270 non-null  float64
 26  location_northing_osgr                          285270 non-null  float64
 27  longitude                                       285270 non-null  float64
 28  latitude                                        285270 non-null  float64
 29  police_force                                    285331 non-null  int64
 30  accident_severity                               285331 non-null  int64
 31  number_of_vehicles                              285331 non-null  int64
 32  number_of_casualties                            285331 non-null  int64
 33  date                                            285331 non-null  object
 34  day_of_week                                     285331 non-null  int64
 35  time                                            285298 non-null  object
```

```
 36  local_authority_(district)                          285331 non-null  int64
 37  local_authority_(highway)                           285331 non-null  object
 38  1st_road_class                                      285331 non-null  int64
 39  1st_road_number                                     285331 non-null  int64
 40  road_type                                           285331 non-null  int64
 41  speed_limit                                         285331 non-null  int64
 42  junction_detail                                     285331 non-null  int64
 43  junction_control                                    285331 non-null  int64
 44  2nd_road_class                                      285331 non-null  int64
 45  2nd_road_number                                     285331 non-null  int64
 46  pedestrian_crossing-human_control                   285331 non-null  int64
 47  pedestrian_crossing-physical_facilities             285331 non-null  int64
 48  light_conditions                                    285331 non-null  int64
 49  weather_conditions                                  285331 non-null  int64
 50  road_surface_conditions                             285331 non-null  int64
 51  special_conditions_at_site                          285331 non-null  int64
 52  carriageway_hazards                                 285331 non-null  int64
 53  urban_or_rural_area                                 285331 non-null  int64
 54  did_police_officer_attend_scene_of_accident         285331 non-null  int64
 55  lsoa_of_accident_location                           268252 non-null  object
 56  casualty_reference                                  186072 non-null  float64
 57  casualty_class                                      186072 non-null  float64
 58  sex_of_casualty                                     186072 non-null  float64
 59  age_of_casualty                                     186072 non-null  float64
 60  age_band_of_casualty                                186072 non-null  float64
 61  casualty_severity                                   186072 non-null  float64
 62  pedestrian_location                                 186072 non-null  float64
 63  pedestrian_movement                                 186072 non-null  float64
 64  car_passenger                                       186072 non-null  float64
 65  bus_or_coach_passenger                              186072 non-null  float64
 66  pedestrian_road_maintenance_worker                  186072 non-null  float64
 67  casualty_type                                       186072 non-null  float64
 68  casualty_home_area_type                             186072 non-null  float64
 69  casualty_imd_decile                                 186072 non-null  float64
dtypes: float64(18), int64(47), object(5)
```

Figure 4.1.1: Data Overview

The dataset's descriptive statistics, which were generated using the describe() method, are shown in Figure 4.1.2. In addition to providing summary statistics like count, mean, standard deviation, and five-number summary, it delivered the description of the data in the DataFrame. Observing the outcome, it was evident that:

- vehicle_manoeuvre: The analysis of vehicle manoeuvre indicated that the mean value was 13.45, with a standard deviation of 5.00. This suggested a diverse range of manoeuvres observed during road accidents.
- speed_limit: The average speed limit in the dataset was 39.23, with a standard deviation of 14.43. This information provided insights into the prevailing speed conditions during recorded accidents.
- weather_conditions: The mean value for weather conditions was 1.49, with a standard deviation of 1.49. This implied a tendency towards clear weather conditions, although the variability suggested the presence of diverse weather situations.

- road_surface_condition: The road surface conditions attribute revealed a mean value of 1.29 and a standard deviation of 0.54. This pointed to generally stable road surfaces, with limited variability observed.
- age_of_driver: The age of drivers in the dataset was characterised by a mean value of 37.42 and a standard deviation of 17.99. This indicated a moderately diverse age distribution among individuals involved in accidents.
- vehicle_reference: The analysis of vehicle reference unveiled a mean value of 1.50, with a standard deviation of 0.66. This suggested a relatively consistent representation of various vehicle types in the dataset.
- casualty_severity: Examining casualty severity revealed a mean value of 2.86, with a standard deviation of 0.37. This highlighted a predominance of accidents with a moderate severity level, as depicted by the dataset.
- vehicle_imd_conditions: The vehicle IMD decile attribute showcased a uniform distribution with a mean value of -1. This suggested that the IMD decile information might not have been fully available or applicable to the dataset.

	vehicle_manoeuvre	speed_limit	weather_conditions	road_surface_conditions
count	174538.000000	174538.000000	174538.000000	174538.000000
mean	13.454738	39.230540	1.493050	1.293655
std	5.986592	14.427203	1.486486	0.544143
min	-1.000000	0.000000	1.000000	-1.000000
25%	9.000000	30.000000	1.000000	1.000000
50%	18.000000	30.000000	1.000000	1.000000
75%	18.000000	50.000000	1.000000	2.000000
max	18.000000	70.000000	9.000000	5.000000

	age_of_driver	vehicle_reference	casualty_severity	vehicle_imd_decile
count	174538.000000	174538.000000	174538.000000	174538.0
mean	37.417938	1.495812	2.864717	-1.0
std	17.986832	0.661851	0.367308	0.0
min	-1.000000	1.000000	1.000000	-1.0
25%	24.000000	1.000000	3.000000	-1.0
50%	35.000000	1.000000	3.000000	-1.0
75%	49.000000	2.000000	3.000000	-1.0
max	97.000000	32.000000	3.000000	-1.0

Figure 4.1.2: Descriptive statistics of Data

Figure 4.1.3 displayed the total number of data in the dataset. The number of rows and columns in the data were determined using shape(). It revealed that there were 70 columns of attributes and 285,331 rows of data.

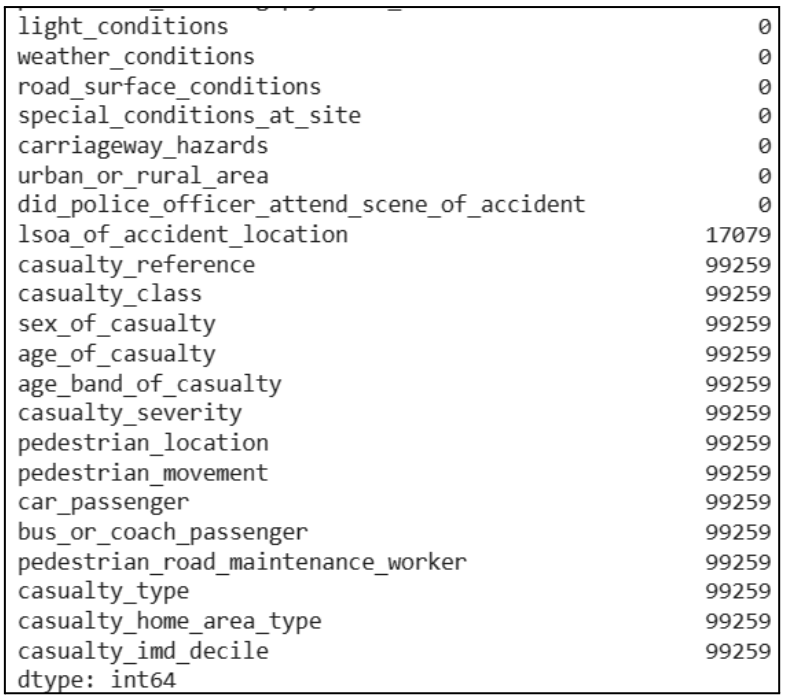
```
Dataset size:

Number of rows: 285331
Number of columns: 70
```

4.1.3: Shape of Data

Figure 4.1.4 showed the missing values in the dataset using isna(). It provided a representation of the dataset's missing values and data types among data attributes. A zero value indicated the absence of any missing values in that specific column. It was determined that there were 61 missing values in location_easting_osgr, location_northing_osgr, longitude, and latitude, 33 missing values in time, 17,079 missing values in lsoa_of_accident_location, and 99,259 missing values in casualty_reference, casualty_class, sex_of_casualty, age_of_casualty, age_band_of_casualty, casualty_severity, pedestrian_location, pedestrian_movement, car_passenger, bus_or_coach_passenger, pedestrian_road_maintenance_worker, casualty_type, casualty_home_area_type, and casualty_imd_decile.

Instead of changing the missing values to mean or mode, it was decided to remove the missing values using dropna() to avoid imputation assumptions and maintain statistical validity, and to simplify the analysis. This decision is shown in Figure 4.1.5.

journey_purpose_of_driver	0
sex_of_driver	0
age_of_driver	0
age_band_of_driver	0
engine_capacity_(cc)	0
propulsion_code	0
age_of_vehicle	0
driver_imd_decile	0
driver_home_area_type	0
vehicle_imd_decile	0
NUmber_of_Casualities_unique_to_accident_index	0
No_of_Vehicles_involved_unique_to_accident_index	0
location_easting_osgr	61
location_northing_osgr	61
longitude	61
latitude	61
police_force	0
accident_severity	0
number_of_vehicles	0
number_of_casualties	0
date	0
day_of_week	0
time	33
local_authority_(district)	0
local_authority_(highway)	0
1st_road_class	0
1st_road_number	0
road_type	0
speed_limit	0
junction_detail	0
junction_control	0
2nd_road_class	0
2nd_road_number	0
pedestrian_crossing-human_control	0
pedestrian_crossing-physical_facilities	0



```
light_conditions                                0
weather_conditions                              0
road_surface_conditions                         0
special_conditions_at_site                      0
carriageway_hazards                             0
urban_or_rural_area                             0
did_police_officer_attend_scene_of_accident     0
lsoa_of_accident_location                   17079
casualty_reference                          99259
casualty_class                              99259
sex_of_casualty                             99259
age_of_casualty                             99259
age_band_of_casualty                        99259
casualty_severity                           99259
pedestrian_location                         99259
pedestrian_movement                         99259
car_passenger                               99259
bus_or_coach_passenger                      99259
pedestrian_road_maintenance_worker          99259
casualty_type                               99259
casualty_home_area_type                     99259
casualty_imd_decile                         99259
dtype: int64
```

Figure 4.1.4: Missing value

```
accident_index                              0
vehicle_reference                           0
vehicle_type                                0
towing_and_articulation                     0
vehicle_manoeuvre                           0
vehicle_location-restricted_lane            0
junction_location                           0
skidding_and_overturning                    0
hit_object_in_carriageway                   0
vehicle_leaving_carriageway                 0
hit_object_off_carriageway                  0
1st_point_of_impact                         0
was_vehicle_left_hand_drive?                0
journey_purpose_of_driver                   0
sex_of_driver                               0
age_of_driver                               0
age_band_of_driver                          0
engine_capacity_(cc)                        0
propulsion_code                             0
age_of_vehicle                              0
driver_imd_decile                           0
driver_home_area_type                       0
vehicle_imd_decile                          0
```

```
NUmber_of_Casualities_unique_to_accident_index      0
No_of_Vehicles_involved_unique_to_accident_index    0
location_easting_osgr                               0
location_northing_osgr                              0
longitude                                           0
latitude                                            0
police_force                                        0
accident_severity                                   0
number_of_vehicles                                  0
number_of_casualties                                0
date                                                0
day_of_week                                         0
time                                                0
local_authority_(district)                          0
local_authority_(highway)                           0
1st_road_class                                      0
1st_road_number                                     0
road_type                                           0
speed_limit                                         0
junction_detail                                     0
junction_control                                    0
2nd_road_class                                      0
2nd_road_number                                     0
pedestrian_crossing-human_control                   0
pedestrian_crossing-physical_facilities             0
light_conditions                                    0
weather_conditions                                  0
road_surface_conditions                             0
special_conditions_at_site                          0
carriageway_hazards                                 0
urban_or_rural_area                                 0
did_police_officer_attend_scene_of_accident         0
```

```
lsoa_of_accident_location                 0
casualty_reference                        0
casualty_class                            0
sex_of_casualty                           0
age_of_casualty                           0
age_band_of_casualty                      0
casualty_severity                         0
pedestrian_location                       0
pedestrian_movement                       0
car_passenger                             0
bus_or_coach_passenger                    0
pedestrian_road_maintenance_worker        0
casualty_type                             0
casualty_home_area_type                   0
casualty_imd_decile                       0
dtype: int64
```

Figure 4.1.5: Result after Fill missing values

The 'number_of_casualties' values, represented as integers, were categorised into three classes for classification models: low, medium, and high. Specifically, values 1 to 7 were classified as low, values 8 to 16 as medium, and values 17 and above as high. After this conversion, the dataset comprised 173,459 instances in the 'low' class, 840 instances in the medium class, and 239 instances in the high class, as displayed in Figure 4.1.6.

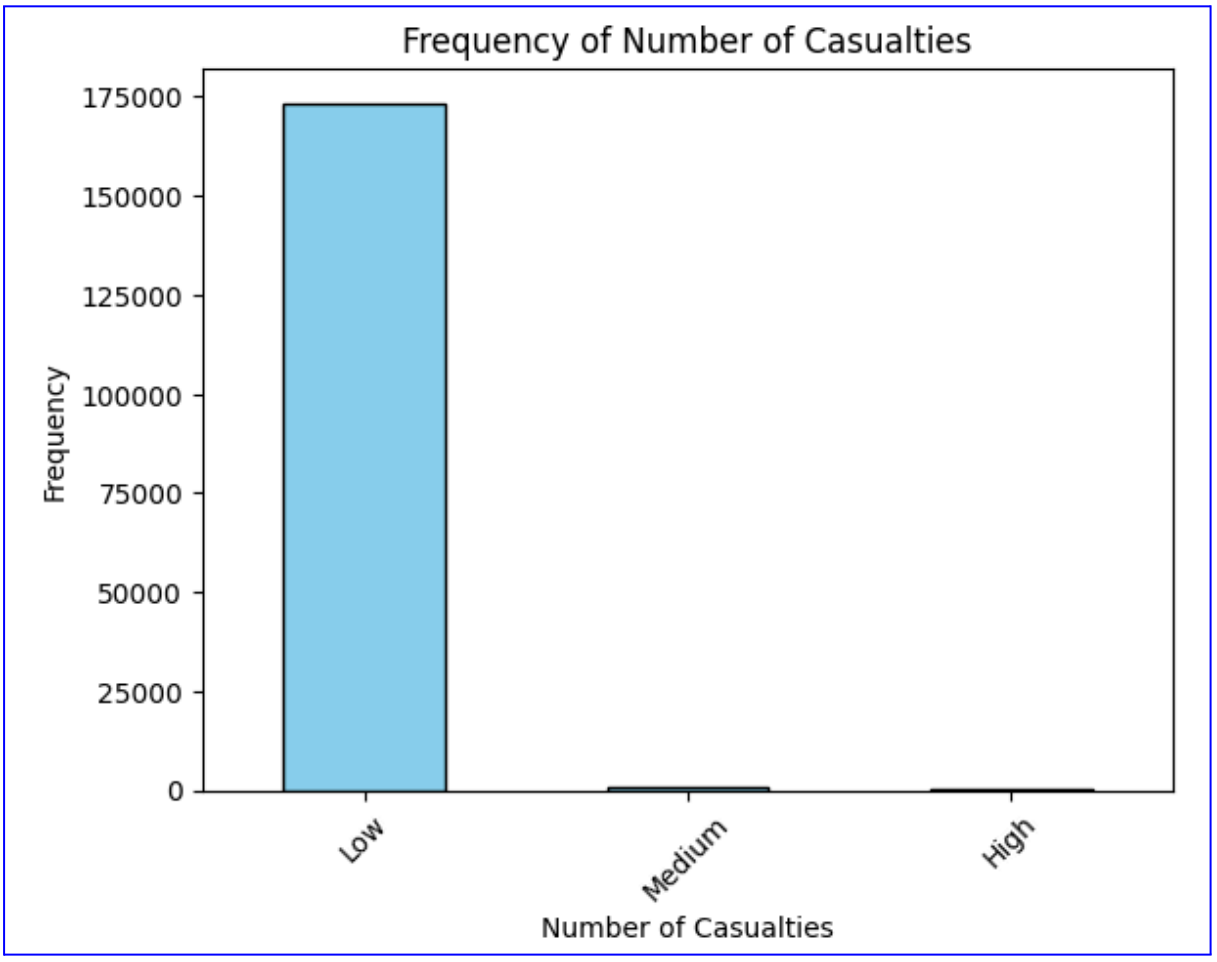


Figure 4.1.6: Frequency of number of casualties

4.2 Exploratory Data Analysis

Figure 4.2.1 displayed a plot chart about the distribution of Accidents by type of vehicles. The x-axis represented 'Vehicle Type', and the y-axis represented the 'Number of accidents', with values starting at 0 and the highest visible bar just above 200,000. The chart showed a significant difference in the number of accidents by vehicle type. With almost 200,000 accidents, the 'Car' vehicle category had by far the largest accident rate. The majority of the other vehicle categories had fewer than 25,000 accidents, and some had even fewer. All other vehicle types had substantially fewer accidents.

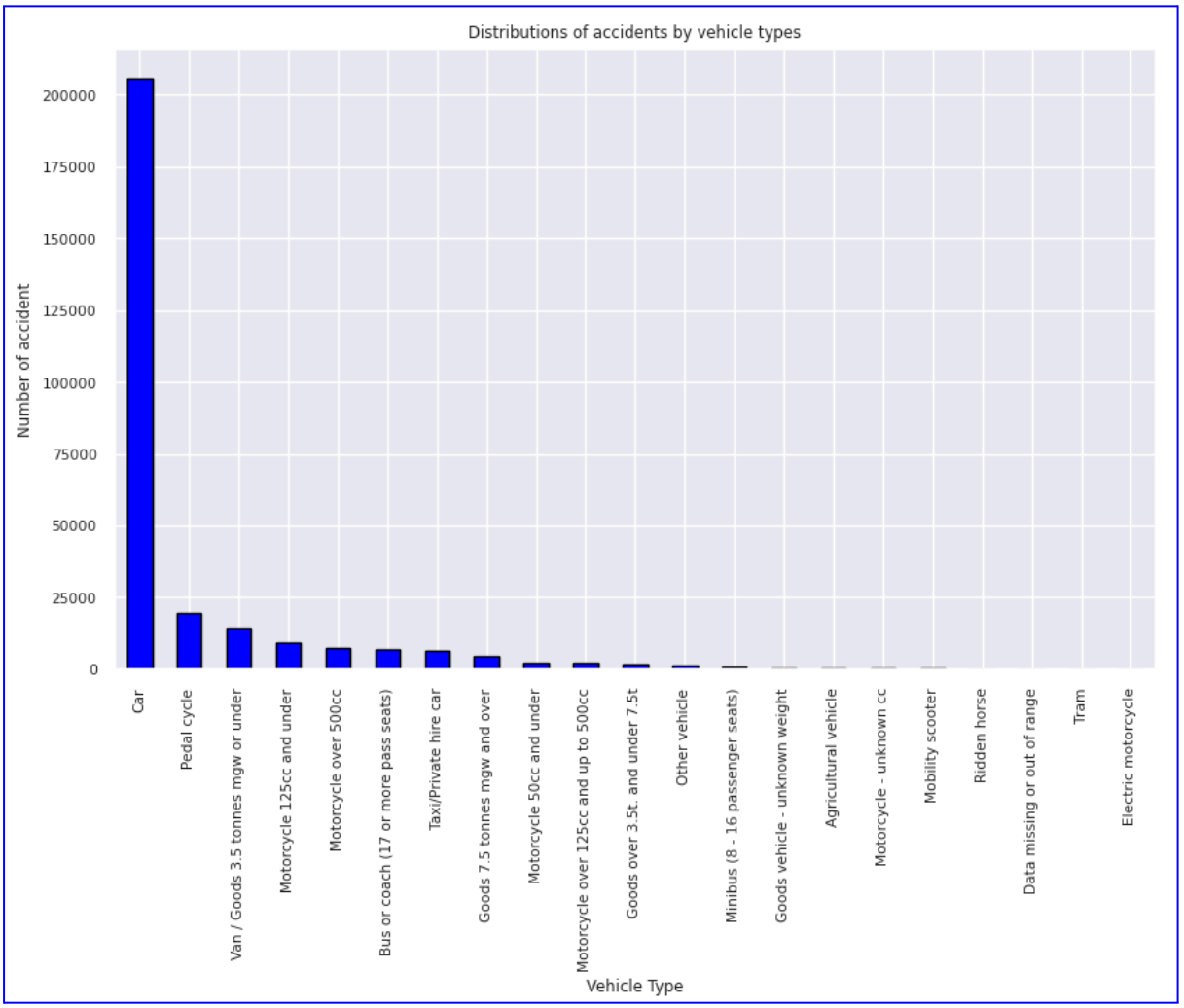


Figure 4.2.1: Distribution of Accidents by Vehicle Type

Figure 4.2.2 displayed a clustered bar chart titled "Accident Severity by Location Type." There were two main categories on the x-axis: Urban and Rural. For each location type, there were three bars representing different levels of accident severity: Slight, Serious, and Fatal, indicated by different colors. In both urban and rural areas, the slight category had the highest number of accidents. For both locations, the serious and fatal categories were much lower than the slight category. The difference in the number of slight accidents between urban and rural areas was substantial, indicating that slight accidents were far more common in urban settings. The fatal category had the lowest count in both areas, which was expected as fatal accidents were less frequent.

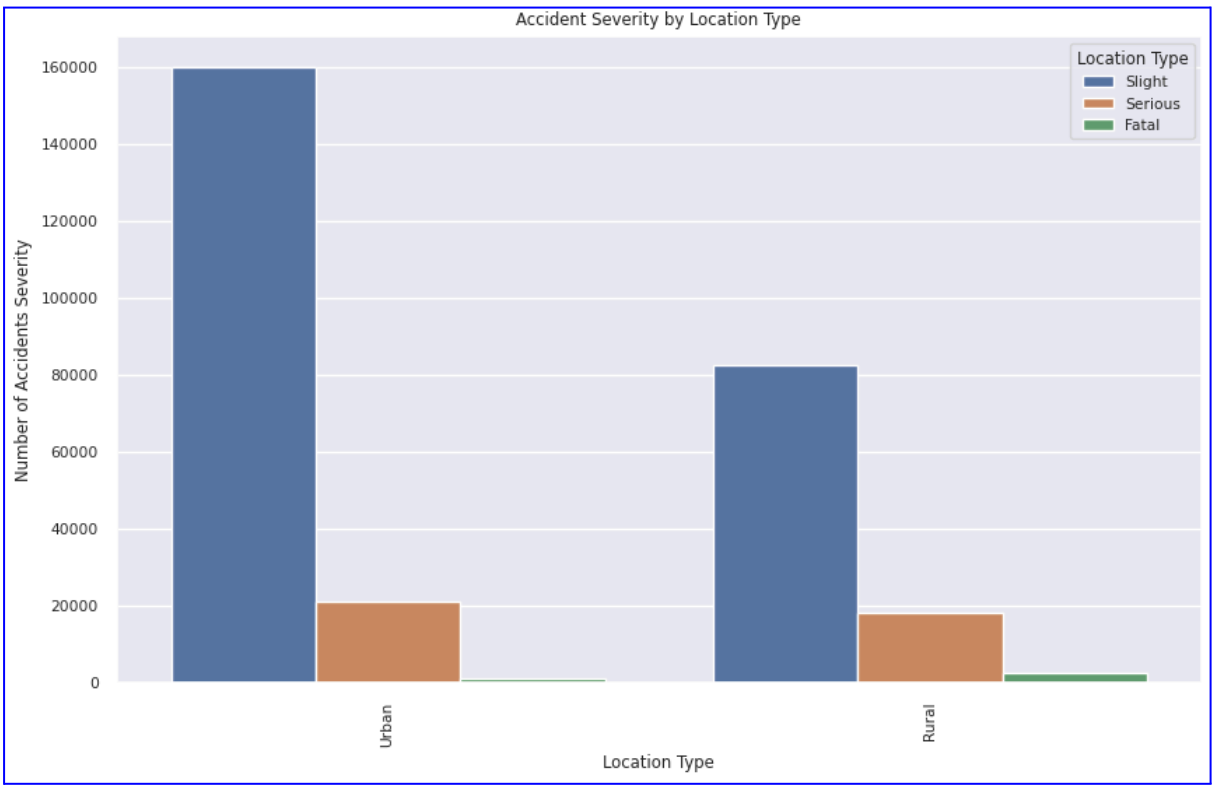


Figure 4.2.2: Accident Severity by Location Type

Figure 4.2.3 displayed a line graph titled "Trend of Accidents Over Time." The x-axis represented time, labeled in a year-month format from January 2015 to January 2016. On the y-axis, the number of incidents ranged from less than 500 to just over 1,000. Throughout the year, there was noticeable variation in the number of accidents. Occasional peaks above 900 incidents and rare troughs below 600 were observed. Rather than a clear upward or downward trend, the graph showed cyclical patterns, suggesting seasonal or cyclical changes. Monthly, there was instability, with sudden jumps and decreases between points. The start and end of the year seemed to have lower points, while the middle and end brought the highest peaks.

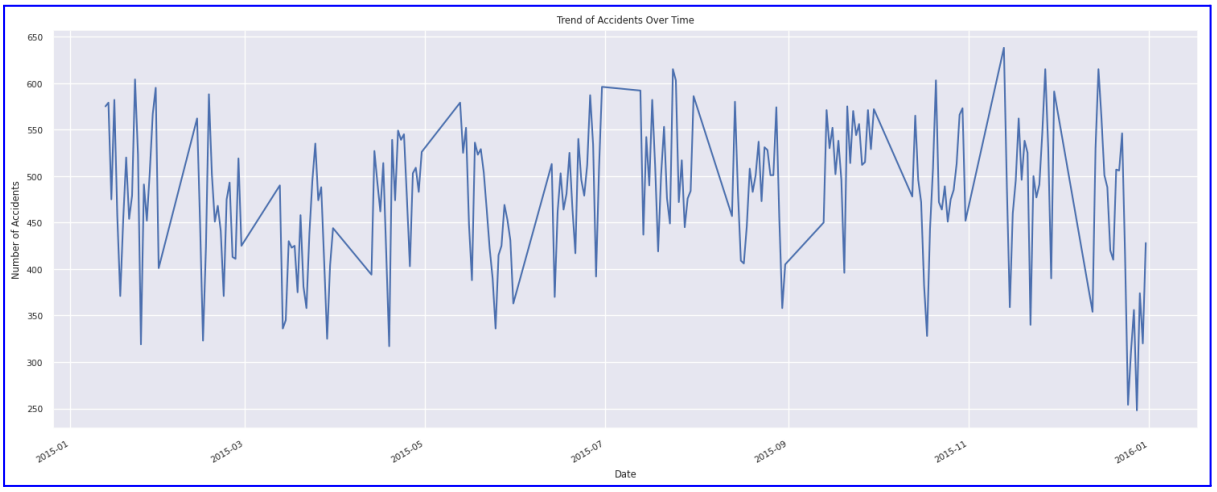


Figure 4.2.3: Trend of Accidents Over Time

Figure 4.2.4 displayed a bar chart titled “Frequency of Age Bands of Casualties”. The x-axis represented different age bands of casualties, while the y-axis represented the frequency of casualties in each age band. The majority of casualties were found in the age group of 26-35, with a noticeable decline in frequency as age increased beyond this range, particularly in the over 75 age group. The 21-25 and 16-20 age brackets also showed relatively high frequencies, highlighting the involvement of younger adults and late teenagers in a significant number of incidents. Conversely, the lowest frequencies were observed in the youngest (0-5) and oldest (over 75) age bands. Additionally, there was a frequency in the "Out of range" category, indicating the presence of data points that didn't fit into the specified age bands.

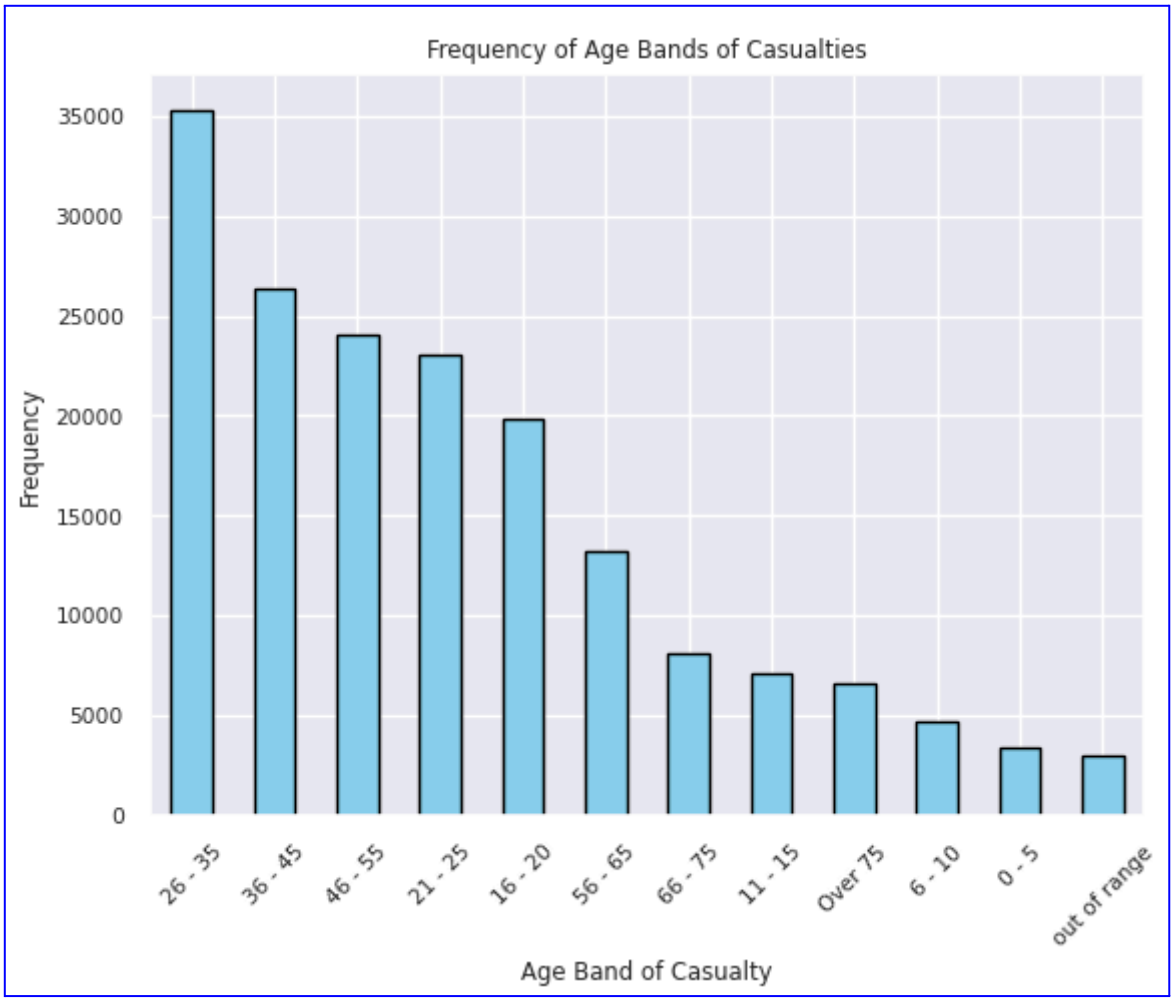


Figure 4.2.4: Distribution of Casualty Age Band

Figure 4.2.5 displayed a correlation matrix heatmap. This map was used to show the correlation coefficient between a larger number of variables. Dark red colors indicated a stronger positive correlation, which was close to 1, while dark blue colors indicated a stronger negative correlation, which was close to -1. Besides that, the diagonal line running from top-left to bottom-right represented individual correlation, which was always 1, and the darkest shade of colour represented positive correlation.

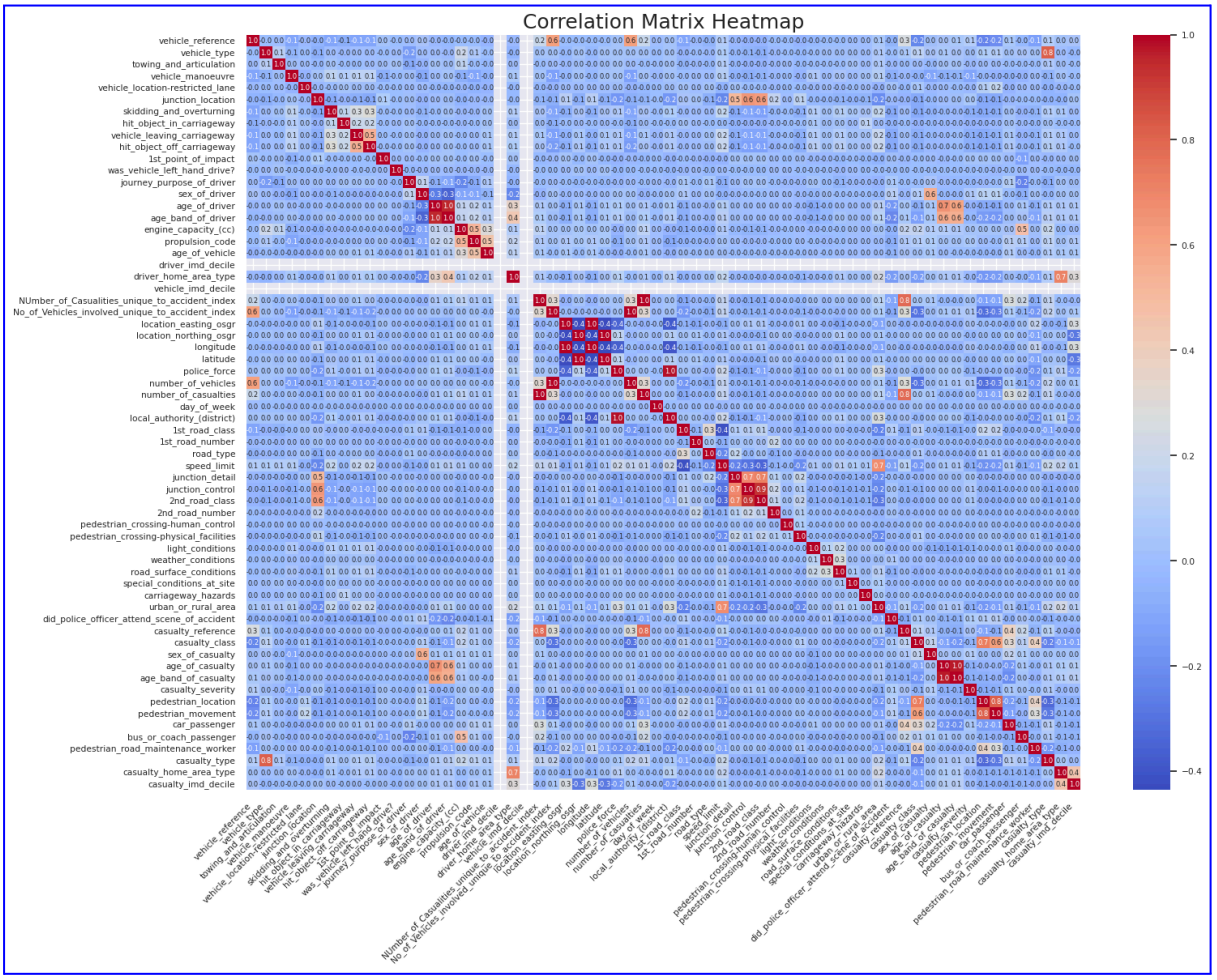


Figure 4.2.5: Heatmap

4.3 Feature Selection

Feature selection was a process that chose an appropriate number of features to include in a model's construction. It could lower the modelling's cost and calculation time while enhancing the model's functionality.

Therefore, in this section, we performed the **Recursive Feature Elimination (RFE)** to choose the features. RFE used machine learning models to automate the process of selecting relevant features by iteratively fitting models, evaluating feature importance, and eliminating the least important features. This was particularly valuable in situations where there were many features, and manual selection or analysis became challenging.

In our dataset, we had 70 attributes and over 250k rows. To streamline the RFE process, we narrowed down the attribute selection to different categories, as shown below.

Location:

- **Attributes**: 'junction_location', 'urban_or_rural_area', 'road_surface_conditions'.
- **Reasoning**: Location-based attributes provide insights into the spatial distribution of incidents. Junction locations can pinpoint high-risk areas, and distinguishing between urban and rural environments can capture changes in road infrastructure and traffic dynamics. Road conditions highlight environmental factors that contribute to accidents, such as wet or icy roads.

Weather

- **Attributes**: 'weather_conditions', 'light_conditions'.
- **Reasoning**: Weather conditions directly affected road safety. Including attributes such as "weather conditions" allowed the model to understand how factors such as rain, snow, or fog affected the severity of an accident. "Light_conditions" further explained visibility issues at different times of day, revealing scenarios where accidents were more likely to occur.

Driver behaviour

- **Attributes**: 'vehicle_manoeuvre', 'skidding_and_overturning', 'journey_purpose_of_driver', ‘sex_of_driver', 'age_band_of_driver', 'driver_imd_decile',

- **Reasoning**: Understanding driver behaviour was critical to predicting accident outcomes. "Vehicle_manoeuvre" provided information about the actions taken by the driver before the accident. "'Skidding_and_overturning'" referred to a situation in which vehicle control was lost. Demographic information ("Sex_of_driver", "Age_band_of_driver", "Driver_imd_decile") helped identify patterns associated with different driver profiles, allowing for more granular analysis.

Vehicle status

- **Attributes**: 'vehicle_type', 'engine capacity(cc)', 'age of vehicle', 'propulsion code'.
- **Reasoning**: Attributes related to the vehicle provided insights into its characteristics. 'Vehicle_type' allowed for differentiation between various vehicle categories, each with its risk profile. 'Engine_capacity_(cc)' and 'Age_of_vehicle' highlighted potential correlations between vehicle power, age, and accident severity. 'Propulsion_code' gave information about the type of propulsion system, offering insights into technological factors.

Time

- **Attributes**: 'day_of_week', 'time'.
- **Reasoning**: The time factor played an important role in the occurrence of accidents. "Day_of_week" captured weekly patterns, identifying weekdays or weekends when accident rates were higher. Time provided the granularity that enabled the model to identify peak times of day and periods of increased risk.

In these five categories, 17 attributes were selected from the datasets to create a balanced and comprehensive representation of the diverse factors influencing road accidents. This ensured that the model considered a wide range of contextual information, leading to a more robust and interpretable feature selection process tailored to predict accident severity and the number of casualties.

Before initiating the RFE process, it was necessary to set the target variables. The objectives for this study were to predict the accident severity and the number of casualties. After determining the target variables, the dataset was divided into two sets: a training set and a testing set, with a ratio of 80% to 20%.

The dataset was normalized before splitting it into training and testing sets in order to speed up the RFE procedure. By ensuring that every feature had a comparable scale, normalization prevented any single feature from controlling the RFE process.

RFE used a machine learning model selection to determine the feature importance. In this instance, Random Forests served as the RFE model. For selecting features, Random Forests were reliable and efficient.

To continue, ten features were selected from the selected attributes by following the procedures shown in Figure 4.3.1. The features generated by the RFE process were presented in Figures 4.3.2 and 4.3.3.

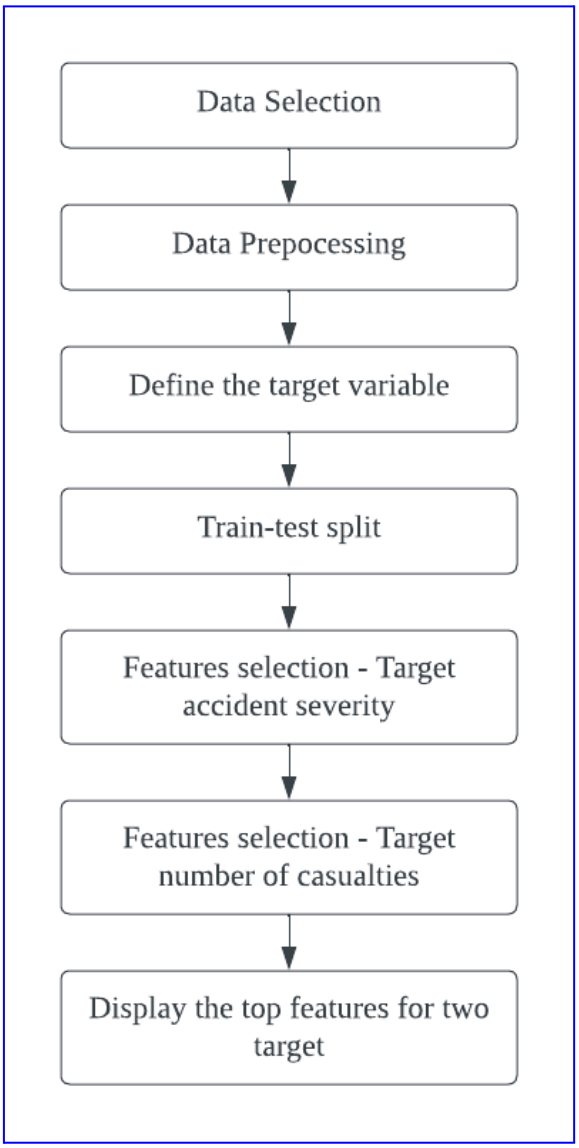


Figure 4.3.1: Feature selection process.

```
Selected Features for Target 1:
Index(['junction_location', 'road_surface_conditions', 'light_conditions',
       'vehicle_manoeuvre', 'skidding_and_overturning',
       'journey_purpose_of_driver', 'age_band_of_driver',
       'engine_capacity_(cc)', 'age_of_vehicle', 'propulsion_code'],
      dtype='object')
```

Figure 4.3.2: Selected Features for accident_severity

```
Selected Features for Target 2:
Index(['junction_location', 'road_surface_conditions', 'light_conditions',
       'vehicle_manoeuvre', 'skidding_and_overturning',
       'journey_purpose_of_driver', 'sex_of_driver', 'age_band_of_driver',
       'engine_capacity_(cc)', 'age_of_vehicle'],
      dtype='object')
```

Figure 4.3.3: Selected features for number_of_casualties

4.5 Data Splitting

Prior to deploying a classification model, the dataset was partitioned into various ratios to ensure the model's generalizability and evaluate its performance across diverse datasets. The datasets were divided into three sets: 80% for training and 20% for testing, 70% for training and 30% for testing, and 60% for training and 40% for testing, as detailed in Table 4.1.

Table 4.1: Training and testing data shape

Train-test split	Training Set: Testing Set
80-20	13930:34908
70-30	122176:52362
60-40	104722:69816

Once the dataset was split, the next steps involved ensuring all the features were on a level playing field through normalisation. This prevented any single feature from dominating the model training process just because of its scale. The issue of imbalanced classes in the classification task was also tackled using SMOTE. This nifty technique generated more examples for the minority class, helping the model learn better and make more accurate predictions for the less common group. Figure 4.5.1 and 4.5.2 displayed examples of data balancing.

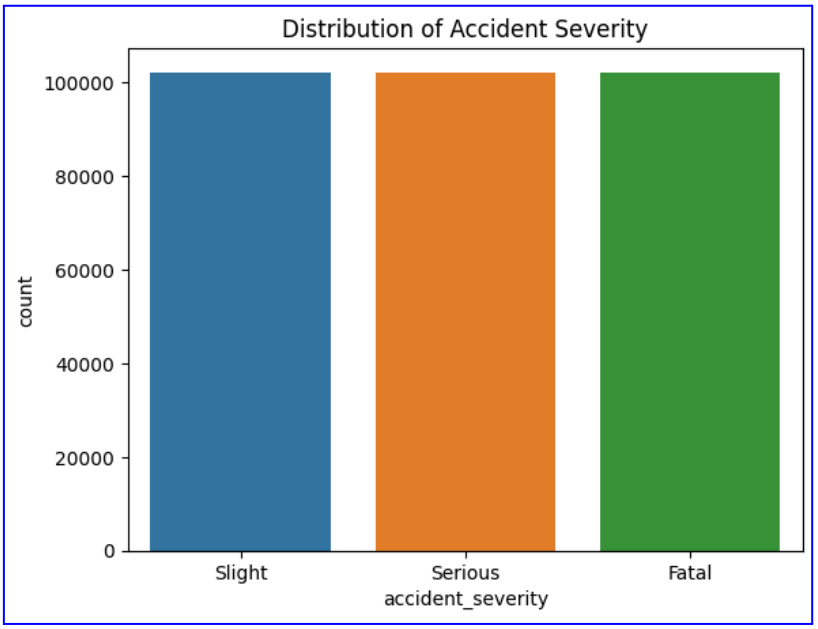


Figure 4.5.1: Distribution of Accident Severity (after data balancing)

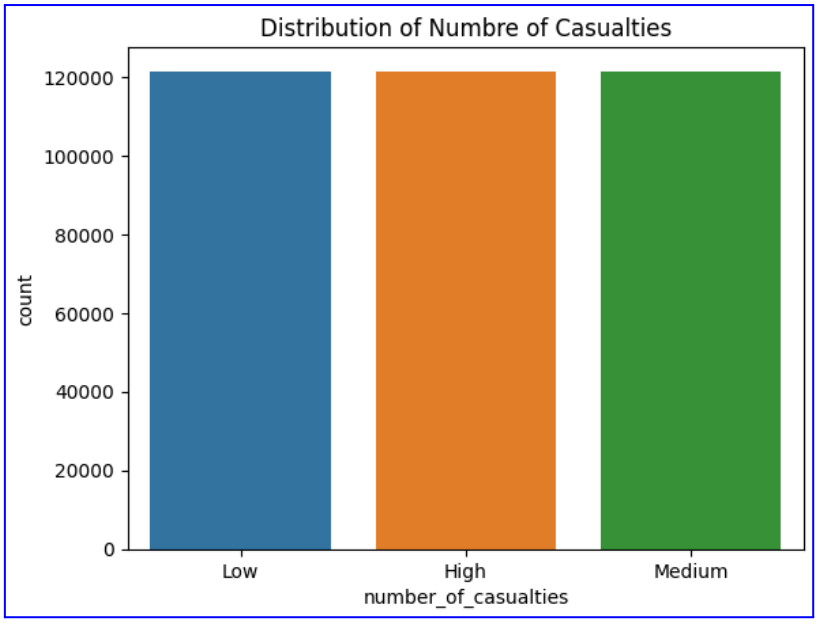


Figure 4.5.1: Distribution of Number of casualties (after data balancing)

5. Predictive Modelling

To investigate the best predictive model, the workflow began with the selection of key attributes that were found in feature selection for predicting both accident severity and the number of casualties in road traffic incidents. These attributes were used to create separate datasets for each prediction task.

Next, the data was divided into two parts for each prediction, one for training the model and another for testing the model. Sometimes, the data might not have had an equal number of examples for each type of accident or severity level. To tackle this issue, a technique called Synthetic Minority Over-sampling Technique (SMOTE) was used. It helped balance out the data so that the model didn't favour one type of prediction over another. This way, the model got a fair chance to learn and make predictions for all situations.

To predict accident severity and the number of casualties, multiple classification models were used to train on the resampled data and make predictions on the test set. The model's performance was then evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. Comparison was made to determine which classification model was the best predictive model. Figure 5.1 showed the flow of the classification process.

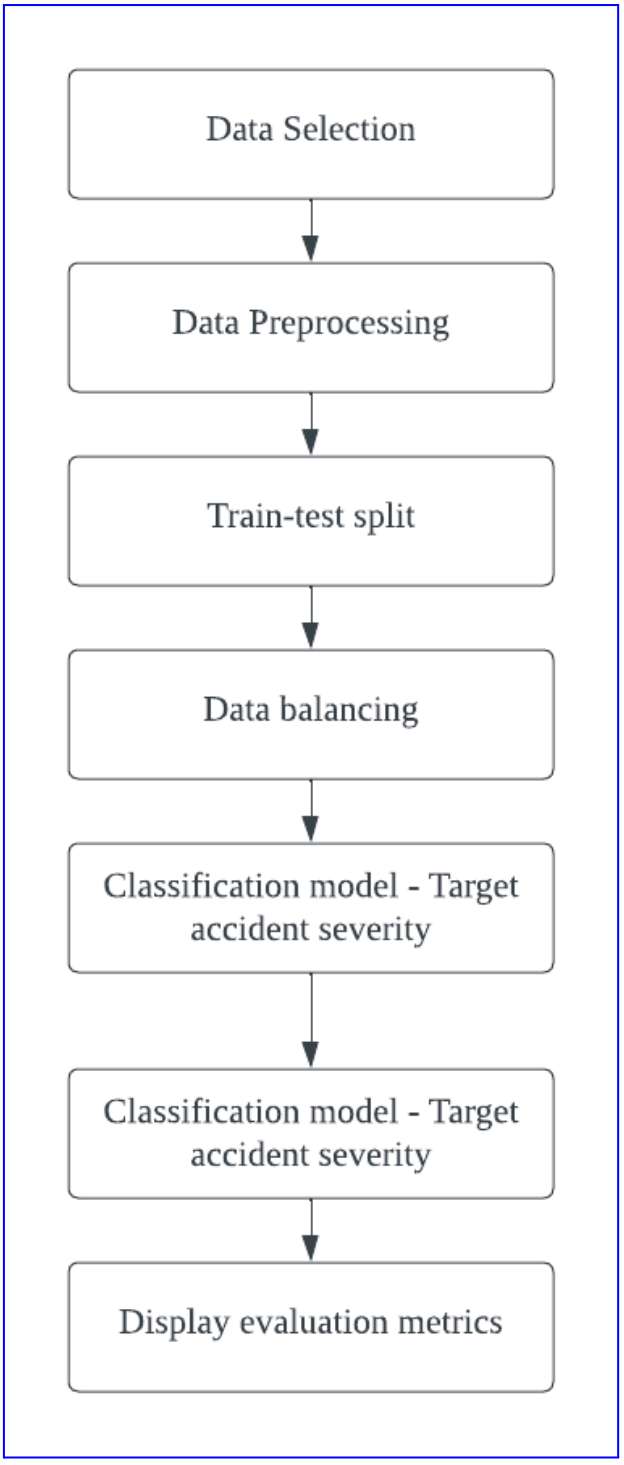


Figure 5.1: Classification process

5.1 Classification

In our project, these methods were selected, namely Random Forest, Logistic Regression, Gradient Boosting, K-Nearest Neighbour (KNN), Naive Bayes, Neural Network, and Decision Tree, to test the dataset and compare their performance.

5.1.1 Random Forest

Random Forest was a powerful technique in computer science where multiple decision trees worked together to make predictions. Each tree independently made decisions based on various factors, and then they all voted to provide a more accurate answer. This approach was robust, flexible, and adept at handling large and complex datasets. Unlike individual decision trees, Random Forest was less prone to making mistakes and was effective in capturing intricate relationships within the data.

5.1.1.1 Classification Result for Random Forest

The random forest model was tested to predict both Accident Severity and Number of Casualties, using different ratios to split the training and testing data, the results are shown in Table 5.1.1.1. For Accident Severity, the 80-20 split showed an accuracy of 0.7169, precision of 0.7945, recall of 0.7169, and F1-score of 0.7477. As moved to a 70-30 split, there were small drops in accuracy to 0.7130 and precision to 0.7890, but recall and F1-score remained pretty steady at 0.7130 and 0.7440 respectively. The 60-40 split led to a further reduction in accuracy to 0.7060 and precision to 0.7810, with recall and F1-score also decreasing slightly to 0.7060 and 0.7370 respectively.

Similar patterns emerged for predicting Number of Casualties. The 80-20 split performed well with an accuracy of 0.9858, precision of 0.9946, recall of 0.9858, and F1-score of 0.9892. The 70-30 split had slight decreases in accuracy to 0.9818 and precision to 0.9944, and the 60-40 split showed slightly lower values across all metrics with accuracy: 0.9821, precision: 0.9946, recall: 0.9821, F1-score: 0.9872. These findings indicate that the model maintained solid performance with different data splits, showing a small trade-off between performance and the amount of training data.

Table 5.1.1.1: Classification Result for Random Forest

		Evaluation Metrics			
Attributes	Train-test split	Accuracy	Precision	Recall	F1-score
Accident Severity	80 Train - 20 Test	0.7169	0.7945	0.7169	0.7477
	70 Train - 30 Test	0.7130	0.7890	0.7130	0.7440
	60 Train - 30 Test	0.7060	0.7810	0.7060	0.7370
Number of Casualties	80 Train - 20 Test	0.9858	0.9946	0.9858	0.9892
	70 Train - 30 Test	0.9818	0.9944	0.9818	0.9869
	60 Train - 30 Test	0.9821	0.9946	0.9821	0.9872

5.1.2 Logistic Regression

Logistic Regression is a type of linear model commonly employed for tasks involving binary classification. Despite its name, it was used to classify, not predict values. It predicted the probability of an instance belonging to a specific class. By applying the logistic function to a combination of input features, it transformed the output into a probability score. Logistic Regression was known for its simplicity, interpretability, and good performance when the connection between features and the target variable was roughly linear.

5.1.2.1 Classification Result for Logistic Regression

The logistic regression model was assessed for its predictive performance in determining Accident Severity and Number of Casualties, employing different train-test split ratios which were shown in Table 5.1.2.1. In the case of Accident Severity, the 70-30 split yielded the highest accuracy at 0.4663 and precision at 0.7696, while the 80-20 split achieved the highest recall at 0.4154 and F1-score at 0.5199. Notably, the 60-40 split exhibited a decrease in all metrics, suggesting a nuanced relationship between data partitioning and model efficacy.

For predicting the Number of Casualties, the 80-20 split showcased the overall best performance with accuracy at 0.6880, precision at 0.9902, recall at 0.6880, and F1-score at 0.8100. These findings illuminated the varying impacts of different train-test splits on the logistic regression model's ability to predict accident outcomes, emphasising the importance of carefully selecting the appropriate data partitioning strategy.

Table 5.1.2.1: Classification Result for Logistic Regression

		Evaluation Metrics			
Attributes	Train-test split	Accuracy	Precision	Recall	F1-score
Accident Severity	80 Train - 20 Test	0.4154	0.7716	0.4154	0.5199
	70 Train - 30 Test	0.4663	0.7696	0.4625	0.5695
	60 Train - 30 Test	0.3998	0.7711	0.3998	0.5111
Number of Casualties	80 Train - 20 Test	0.6880	0.9902	0.6880	0.8100
	70 Train - 30 Test	0.6330	0.9910	0.6330	0.7703
	60 Train - 30 Test	0.6586	0.9914	0.6586	0.7895

5.1.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and direct classification method. It figures out the category of a data point by checking the most common class among its nearest neighbors in the feature space. KNN doesn't assume anything about how the data is spread out. It's important to keep in mind that KNN is influenced by two main factors: the type of distance used to measure closeness and the value of k, which is the number of neighbors considered. Therefore, making careful choices when setting these parameters is essential.

5.1.3.1 Classification Result for K-Nearest Neighbour

The K-Nearest Neighbors (KNN) model was tested to see how well it predicted Accident Severity and Number of Casualties, using different ways of splitting the training and testing data which are shown in Table 5.1.3.1. When it came to Accident Severity, the model consistently showed accuracy, like 0.7151 for the 80-20 split, 0.7112 for the 70-30 split, and 0.7008 for the 60-40 split. Precision and recall, which tell us about the model's exactness and completeness, also stayed pretty stable across the different splits, with F1-scores ranging from 0.7299 to 0.7405.

Now, for predicting the Number of Casualties, the KNN model did really well, always getting high accuracy, precision, recall, and F1-scores. For example, with the 80-20 split, it achieved an accuracy of 0.9818, precision of 0.9935, recall of 0.9818, and an F1-score of 0.9866. These findings highlight that the KNN model is reliable in predicting accident outcomes, especially when it comes to estimating the Number of Casualties. It worked well across different ways of splitting the data for training and testing.

Table 5.1.3.1: Classification Result for K-Nearest Neighbors

		Evaluation Metrics			
Attributes	Train-test split	Accuracy	Precision	Recall	F1-score
Accident Severity	80 Train - 20 Test	0.7151	0.7747	0.7151	0.7405
	70 Train - 30 Test	0.7112	0.7736	0.7112	0.7377
	60 Train - 30 Test	0.7008	0.7696	0.7009	0.7299
Number of Casualties	80 Train - 20 Test	0.9818	0.9935	0.9818	0.9866
	70 Train - 30 Test	0.9802	0.9936	0.9802	0.9858
	60 Train - 30 Test	0.9767	0.9934	0.9767	0.9839

5.1.4 Decision Tree

Decision Trees are straightforward yet effective models for classifying things. They repeatedly divide the data using different features, forming a tree structure. Each endpoint, or leaf, on the tree represents a specific category. Decision Trees are easy to understand and visualise, but there's a risk of overfitting, especially when the tree becomes deep. To prevent this, techniques like pruning can be used to ensure the model generalises well.

5.1.4.1 Classification Result for Decision Tree

The Decision Tree model was performed in predicting Accident Severity and Number of Casualties using different ways of splitting the training and testing data which are shown in Table 5.1.4.1. For Accident Severity, the model consistently showed accuracy, like 0.6707 for the 80-20 split, 0.6705 for the 70-30 split, and 0.6631 for the 60-40 split. Precision, which tells us how exact the predictions are, and recall, indicating how well the model captures all the important instances, had similar patterns across splits, giving F1-scores ranging from 0.7072 to 0.7146.

Now, when it comes to predicting the Number of Casualties, the Decision Tree model always had high accuracy, precision, recall, and F1-scores, such as 0.9827 accuracy, 0.9944 precision, 0.9827 recall, and 0.9873 F1-score for the 80-20 split. These results show that the Decision Tree model is reliable and effective in predicting accident outcomes, working well with different ways of splitting the data for training and testing.

Table 5.1.4.1: Classification Result for Decision Tree

		Evaluation Metrics			
Attributes	Train-test split	Accuracy	Precision	Recall	F1-score
Accident Severity	80 Train - 20 Test	0.6707	0.7918	0.6707	0.7146
	70 Train - 30 Test	0.6705	0.7857	0.6705	0.7138
	60 Train - 30 Test	0.6631	0.7789	0.6631	0.7072
Number of Casualties	80 Train - 20 Test	0.9827	0.9944	0.9827	0.9873
	70 Train - 30 Test	0.9790	0.9942	0.9790	0.9853
	60 Train - 30 Test	0.9784	0.9943	0.9784	0.9851

6. Classification Result Discussion

In evaluating the performance of four different models—Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree—for predicting Accident Severity and Number of Casualties with different train-test split ratios, some key observations emerged.

The Random Forest model consistently showed strong performance across different splits. For Accident Severity, it had accuracy, precision, recall, and F1-scores ranging from 0.7060 to 0.7169. Similarly, when predicting the Number of Casualties, it achieved high accuracy (0.9767 to 0.9858), precision (0.9934 to 0.9946), recall (0.9767 to 0.9858), and F1-scores (0.9839 to 0.9892). This model proved to be reliable, especially excelling in predicting the Number of Casualties.

In contrast, Logistic Regression showed sensitivity to different data splits, with noticeable changes in performance. For Accident Severity, the 70-30 split had the highest accuracy (0.4663) and precision (0.7696), while the 80-20 split achieved the highest recall (0.4154) and F1-score (0.5199). When predicting the Number of Casualties, the 80-20 split stood out with high accuracy (0.6880), precision (0.9902), recall (0.6880), and F1-score (0.8100). However, this model exhibited some variability across splits, indicating a nuanced relationship with data partitioning.

The KNN model consistently demonstrated accuracy and stability in predicting both Accident Severity and Number of Casualties. Across different splits, accuracy ranged from 0.7008 to 0.7151 for Accident Severity and from 0.9818 to 0.9866 for the Number of Casualties. F1-scores ranged from 0.7299 to 0.7405 for Accident Severity and 0.9818 to 0.9866 for the Number of Casualties, showing reliability and effectiveness, particularly in predicting the Number of Casualties.

Similar to the other models, the Decision Tree showed consistent performance for both Accident Severity and Number of Casualties. Accuracy ranged from 0.6631 to 0.6707 for Accident Severity and from 0.9784 to 0.9827 for the Number of Casualties. Precision, recall, and F1-scores also remained stable across different splits. The Decision Tree consistently performed well, with high F1-scores indicating a good balance between precision and recall.

Taking overall performance into account, the Random Forest model seems to be the most reliable for predicting both Accident Severity and Number of Casualties. It consistently demonstrated high accuracy, precision, recall, and F1-scores across different data splits. The ensemble approach of the Random Forest model, combining predictions from multiple decision trees, likely contributed to its stability and effectiveness. This model outperformed Logistic Regression, KNN, and Decision Tree, showcasing its versatility in handling varying data distributions and demonstrating superiority in predicting accident outcomes.

7. Conclusion

In summary, this research explored four different machine learning models—Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree—to predict Accident Severity and Number of Casualties, considering various ways of splitting the training and testing data. Each model showed unique strengths and considerations when dealing with the complexities of our dataset.

The Random Forest model consistently stood out as a strong and reliable performer, achieving high accuracy, precision, recall, and F1-scores for both Accident Severity and Number of Casualties across different data splits. Its approach of combining predictions from multiple decision trees contributed to its stability and effectiveness in predicting accident outcomes.

Logistic Regression, though sensitive to how we split the data, performed well, especially in predicting the Number of Casualties with high precision and accuracy in the 80-20 split. However, its performance varied across different splits, suggesting a complex relationship between how we split the data and the model's effectiveness.

K-Nearest Neighbors (KNN) consistently showed stability and effectiveness in predicting both Accident Severity and Number of Casualties, achieving high accuracy and F1-scores across different splits. This model demonstrated reliability, particularly in estimating the Number of Casualties, making it a strong choice for certain situations.

The Decision Tree model consistently maintained steady performance across splits, achieving good accuracy, precision, recall, and F1-scores. Its ability to strike a balance between precision and recall suggests its potential for predicting accident outcomes effectively.

In conclusion, the choice of the best model depends on the specific goals and requirements of the prediction task. The Random Forest model's strong overall performance makes it a compelling option for making robust and reliable predictions in the context of accident severity and casualty numbers. As the field of machine learning progresses, further exploration and improvements to these models can enhance their ability to predict accident outcomes more effectively.

References

Comi, A., Polimeni, A., & Balsamo, C. (2022). Road Accident Analysis with Data Mining Approach: evidence from Rome. *Transportation Research Procedia*, *62*, 798–805. https://doi.org/10.1016/j.trpro.2022.02.099

Priya, S., & Agalya, R. (2018). Association Rule Mining Approach to Analyze Road Accident Data. *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, 1–5. https://doi.org/10.1109/ICCTCT.2018.8550950

Road Accidents Incidence. (n.d.). Retrieved January 11, 2024, from https://www.kaggle.com/datasets/akshay4/road-accidents-incidence/data

Yassin, S. S., & Pooja. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences*, *2*(9), 1576. https://doi.org/10.1007/s42452-020-3125-1

Eboli, L., Forciniti, C., & Mazzulla, G. (2020). Factors influencing accident severity: an analysis by road accident type. *Transportation Research Procedia*, *47*, 449–456. https://doi.org/10.1016/j.trpro.2020.03.120

Appendix