Name: **Binto George Babu**

Student Number: **202311509**

Introduction

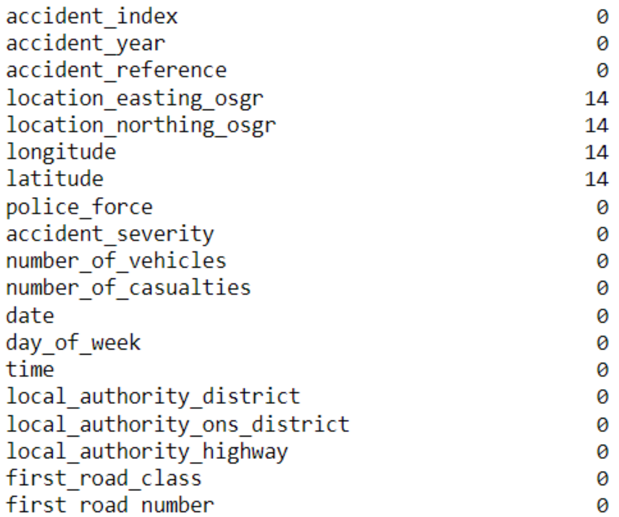
Road traffic accidents are becoming a situation that needs to be handled with the utmost care. To implement ways to mitigate both the occurrence and severity of accidents, it is necessary to understand the dynamics of road traffic accidents first. The primary goal of this report is to offer suggestions to enhance road safety through identifying patterns and risk factors. Finally, a predictive model was developed to enhance the road safety. This report will explore the analysis of data associated with road traffic accidents in 2020. The dataset consists of rich of information that provides valuable insights into various facets of accidents. By utilising data driven methods, the methodology aligns with the objective for creating a safer road environment for everyone.

Methodology

The methodology for this study includes initially retrieving the required data from the accident database. Followed by the data retrieval from the database, to enhance the accuracy, data was pre-processed and cleaned. Exploratory data analysis was conducted for further examining more details regarding the characteristics of the dataset. The apriori algorithm was then employed for pattern mining followed by clustering analysis. To identify any anomalies in the dataset, outlier detection techniques were utilised. Following the feature selection process, the selected features were standardised and then the random forest algorithm was used for predictive analysis to estimate the traffic accident severity.

Data Description and Cleaning

The dataset utilised for this study was Great Britain’s accident database from 2020. There are four distinct tables in the accident database, including accident, vehicle, casualty, and LSOA. The accident and vehicle tables consist of 91,199 and 167,375 entries in 36 and 28 columns respectively. Also, the casualty and lsoa tables consist of 115,584 and 34,378 entries in 19 and 7 columns respectively. The accident table consists of 14 NaN values in the columns including 'longitude', 'latitude', 'location_easting_osgr', and 'location_northing_osgr' (as depicted in Fig. 1). Most of the accidents should be reported and handled by the police units from specific regions. Thus, the missing values were filled by considering the corresponding police stations. This was achieved by

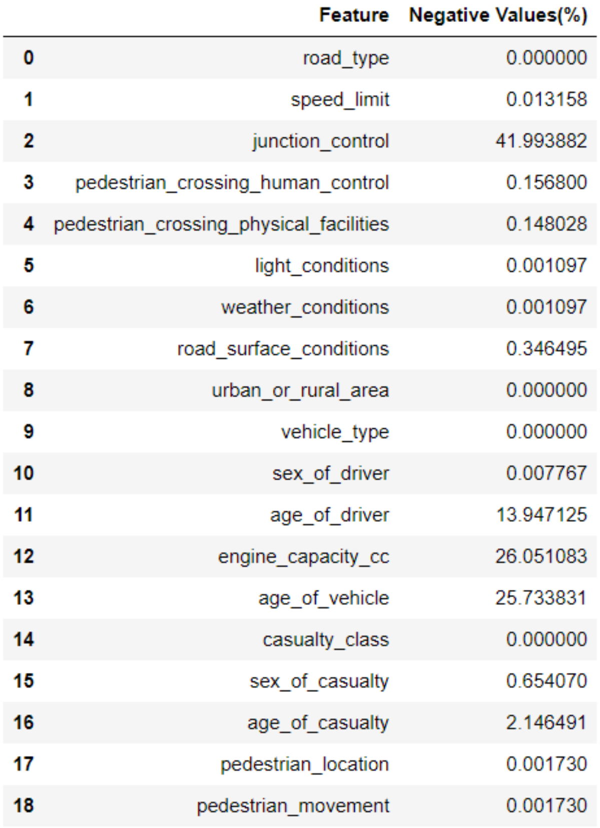


```
accident_index                  0
accident_year                   0
accident_reference              0
location_easting_osgr          14
location_northing_osgr         14
longitude                      14
latitude                       14
police_force                    0
accident_severity               0
number_of_vehicles              0
number_of_casualties            0
date                            0
day_of_week                     0
time                            0
local_authority_district        0
local_authority_ons_district    0
local_authority_highway         0
first_road_class                0
first_road_number               0
```

Fig. 1 NaN values in accident data

considering the location details of corresponding police stations to fill in the missing information in the accident table.

The proportion of negative values in some relevant features across the whole dataset is shown in Fig. 2. To ensure the data integrity and consistency throughout the dataset, the numerical features containing negative values were replaced by the average replacement method. Following the replacement, categorical features containing negative values were subsequently replaced with the mode value.



	Feature	Negative Values(%)
0	road_type	0.000000
1	speed_limit	0.013158
2	junction_control	41.993882
3	pedestrian_crossing_human_control	0.156800
4	pedestrian_crossing_physical_facilities	0.148028
5	light_conditions	0.001097
6	weather_conditions	0.001097
7	road_surface_conditions	0.346495
8	urban_or_rural_area	0.000000
9	vehicle_type	0.000000
10	sex_of_driver	0.007767
11	age_of_driver	13.947125
12	engine_capacity_cc	26.051083
13	age_of_vehicle	25.733831
14	casualty_class	0.000000
15	sex_of_casualty	0.654070
16	age_of_casualty	2.146491
17	pedestrian_location	0.001730
18	pedestrian_movement	0.001730

Fig. 2 Percentage of negative values and corresponding features

In the dataset, the age of the drivers was ranged from 3 to 100. Thus, special attention was taken for cleaning the age data within the dataset. Even though, it's possible to apply for a provisional license at age of 15 years and 9 months old, the legal age for driving in UK is 17 or above (Government Digital Service, 2015). This discrepancy was rectified by replacing the values with the median age between 17 and the oldest age. This approach ensures that the age data remains within the appropriate range. The date column was converted into DateTime format. Moreover, two additional columns, namely 'hours' and 'minutes' were added for further analysis.

Accident Demography

For making data driven decisions to improving road safety, it is vital to understand the demographic distribution of road traffic accidents. The analysis reveals that approximately 78% of accidents fall under the category of minor accidents such as slight accidents (Fig. 3).

These accidents generally cause little damage or injuries. Around 20% of accidents are classified as serious, indicating more severe injuries and damage. Notably, only 1.5% of the total accidents are categorised as severe accidents with fatal injuries.

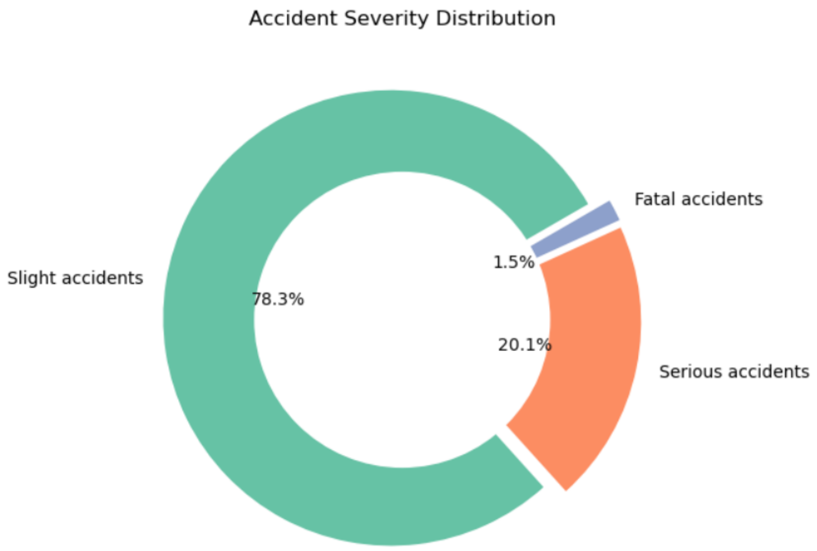


Fig. 3 Accident Severity Distribution

Majority of the drivers involved in the accidents falls under the age range between 21 and 45. In addition to that, comparing with females, there exist a higher proportion of male drivers and male casualties, as shown in Fig. 4 and Fig 5. This implies that those within this age group are more likely to drive and involved in accidents.

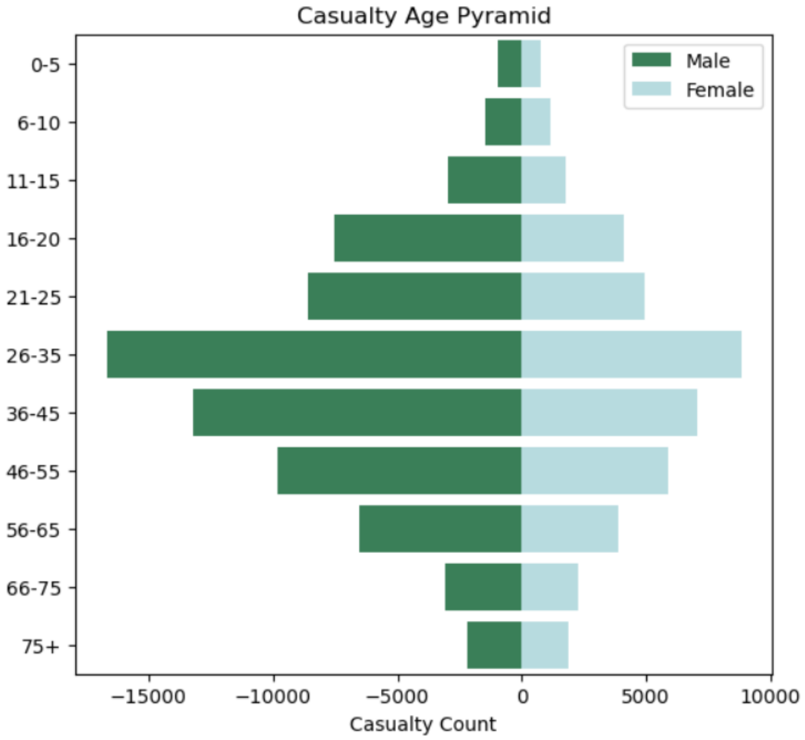


Fig. 4 Drivers Age Pyramid

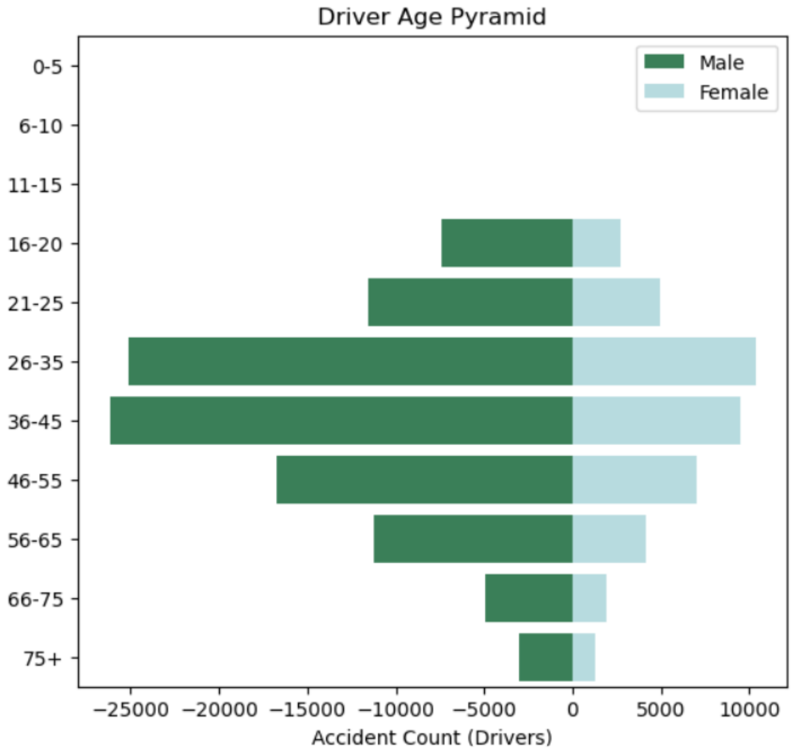


Fig. 5 Casualty Age Pyramid

Data Insights

1. Significant hours of the day and days of the week for accident occurrence.

From the analysis, it's clear that, the accidents occur throughout the day with notable trends observed during certain hours. About 5.8% accidents is observed in the morning around 8 AM. But a higher percentage of accidents have been observed during the hours between 3-5 PM reach its peak at 5 PM. The gradual rise in accidents from 10 AM to its peak at 5 PM indicates peak traffic hours as people commute to and from work (Fig. 6).

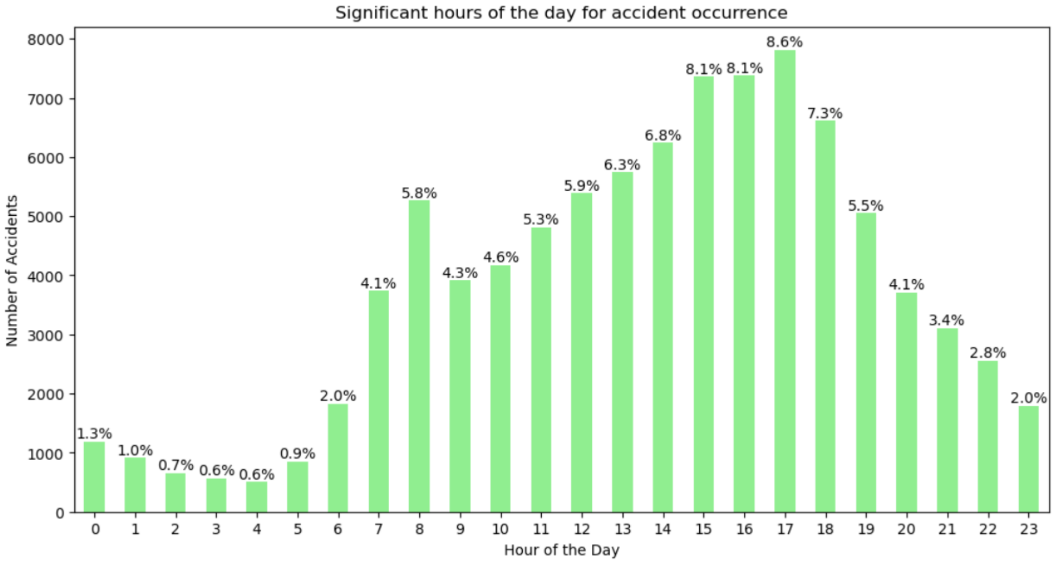


Fig. 6 Significant hours of the day for accident occurrence.

Majority of accidents occur during weekdays, with highest accident occurrence approximately at 16.3% on Fridays. On the other hand, Sundays have a lower proportion of accidents about 11.3%. The spike on Friday resulted from increase in vehicle activity due to people rushing home for the weekend (Fig. 7).

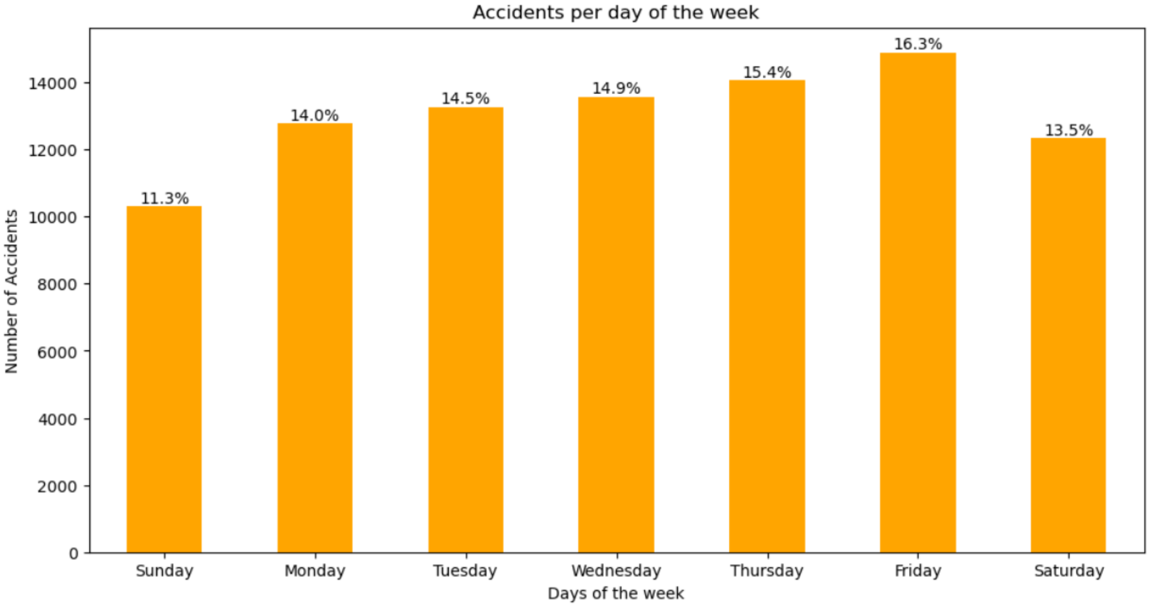


Fig. 7 Accidents per days of the week

2. **Significant hours of the day and days of the week for accident occurrence for motorbikes.**

Most of the motorbike accidents are happened between 2 PM and 6 PM, with a significant peak of 9.9% total accidents occurring around 4 PM (Fig. 8). Even though, there are different categories of motorbikes, those with 50cc to 125cc engines have the highest accident rates. In terms of days of the week, Fridays are the ones that have the highest accident occurrences for 50cc and under, 50cc to 125 cc, and 125cc to 500cc engine motorbikes. However, the day with the highest accident rate that corresponds to motorbikes with 500cc engines are Sundays (Fig. 9).

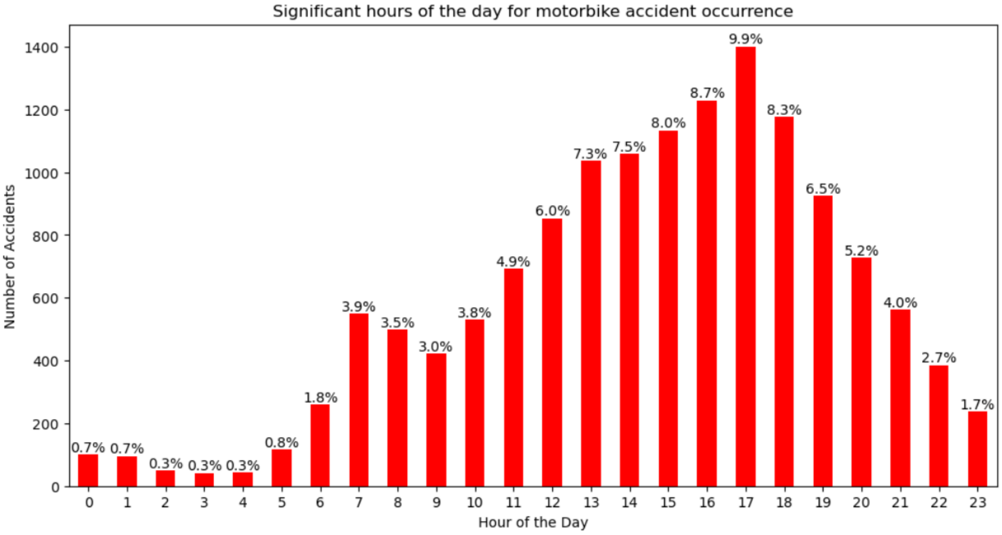


Fig. 8 Hours of the day for the occurrence of motorbike accidents.

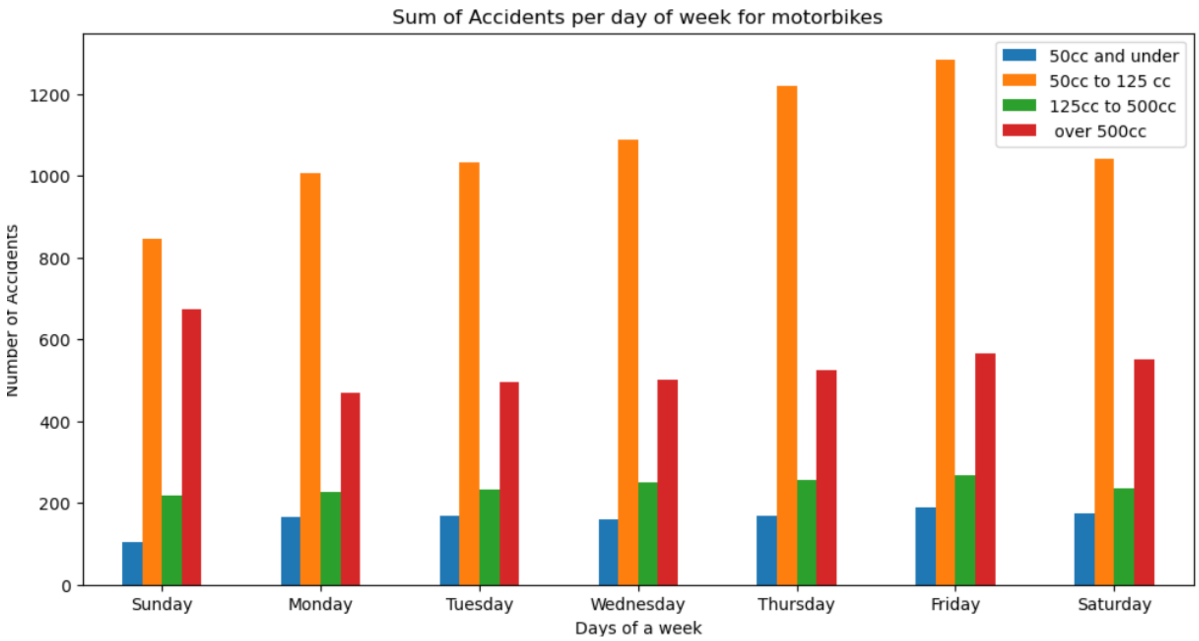


Fig. 9 Total accidents per day of a week corresponds to the motorbikes.

3. Significant hours of the day and days of the week when pedestrian accidents occur.

Higher proportion of pedestrian accidents happened between 3 PM and 6PM of a day, with a peak of 11.3% of total accidents around 3 PM. Furthermore, during the early rush hours, there were notable pedestrian accidents around 8 AM (Fig. 10).

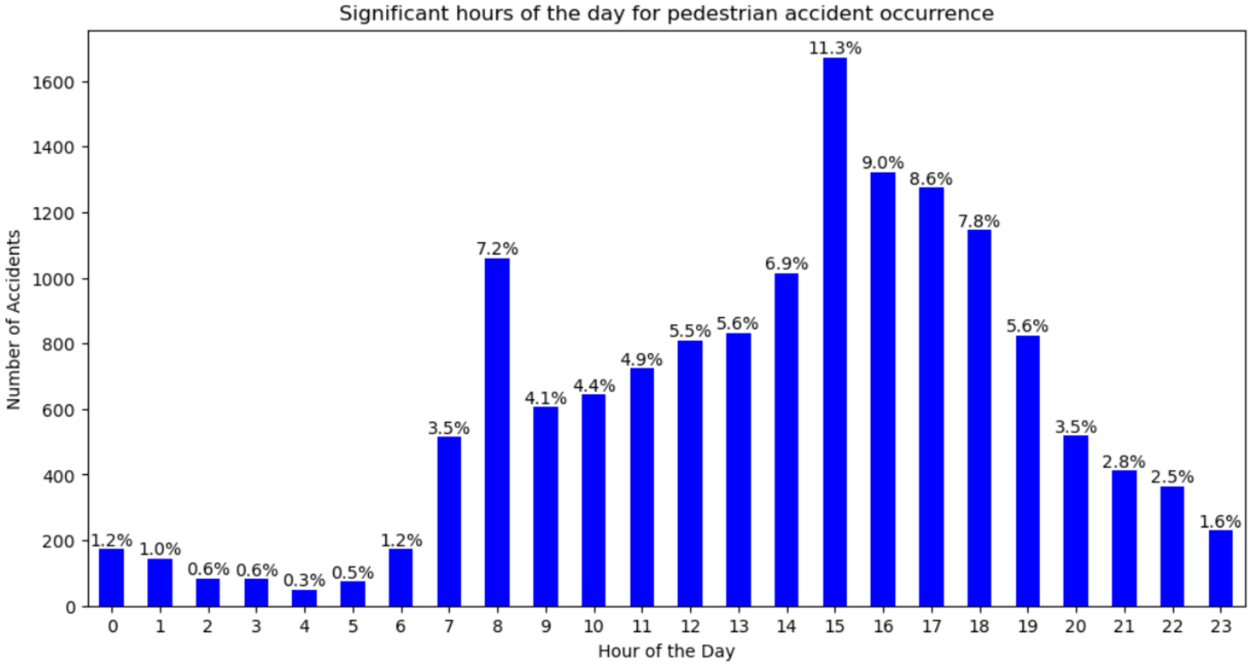


Fig. 10 Hours of the day for the occurrence of pedestrian accidents.

The notable decrease in accidents after 6 PM indicates that most pedestrian accidents occur during the rush hours. Additionally, with a higher peak observed on Fridays, most of the pedestrian accidents are happened throughout the weekdays (Fig. 11). Every day, a higher number of accidents involving drivers or riders occur than that of pedestrian accidents (Fig. 12).

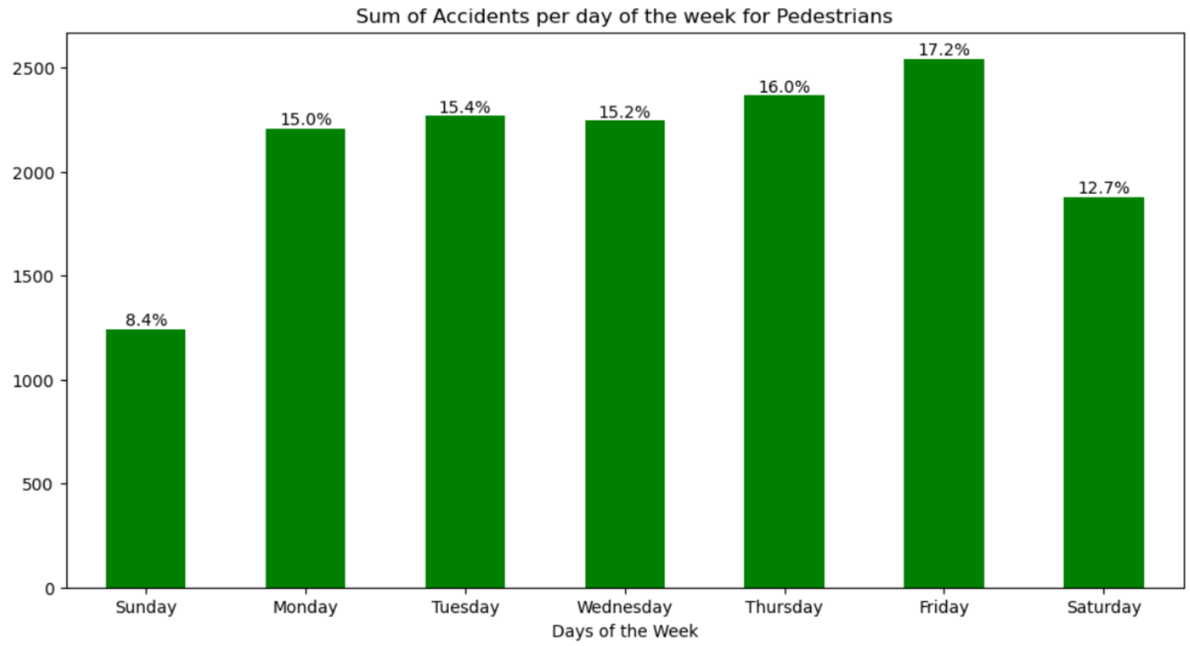


Fig. 11 Total pedestrian accidents per day of the week.

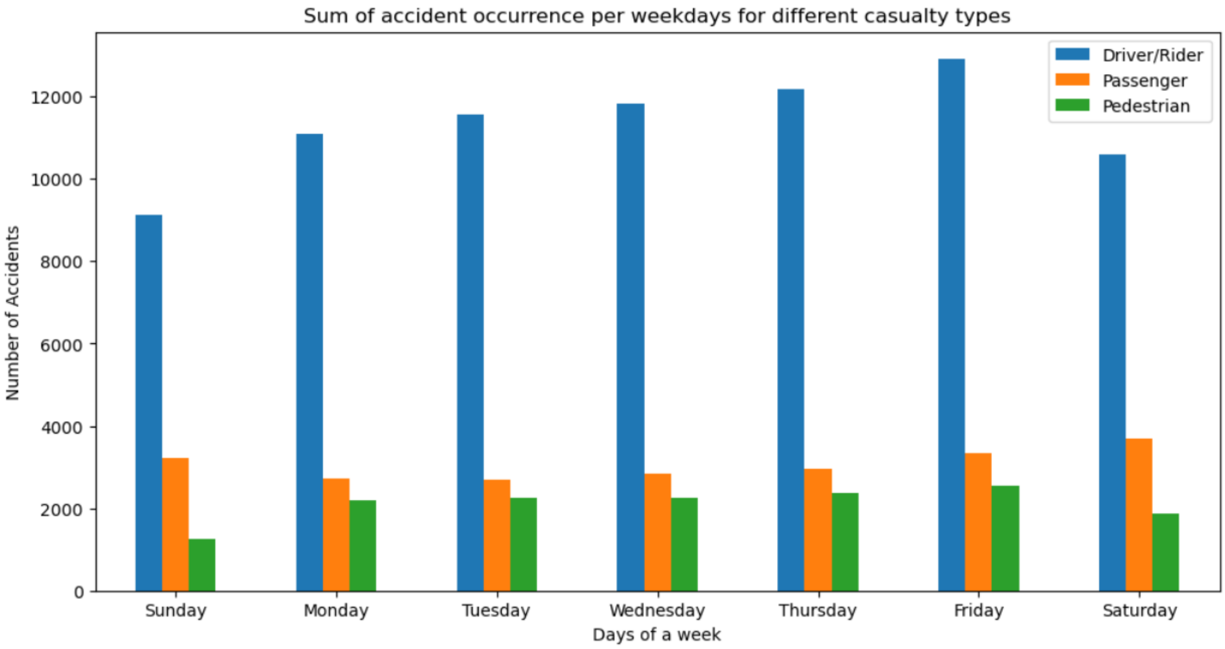


Fig. 12 Total accident occurrence per days of a week for different casualty types.

4. Effect of light and weather condition.

Interestingly, accidents aren't significantly impacted by severe weather. Over 70% of accidents happen during the day in fine weather conditions with no high winds. Darkness and poor weather often lead to a decrease the overall number of accidents as make driving undesirable, resulting in fewer vehicles on the road (Fig. 13 & Fig. 14).

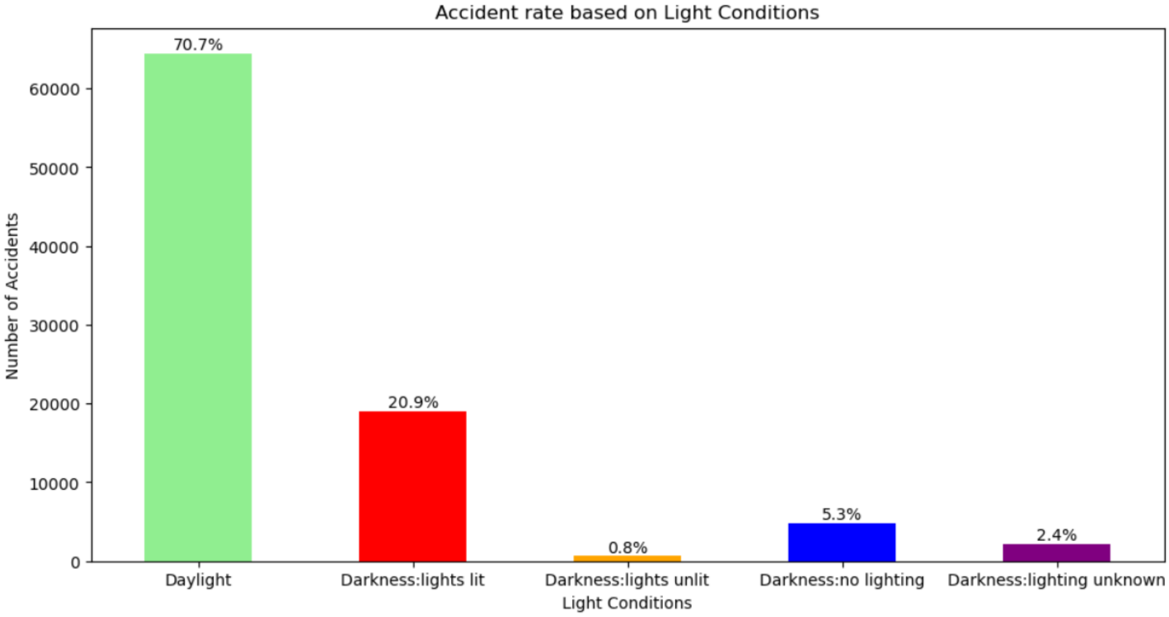


Fig. 13 Effect of light conditions on accident.

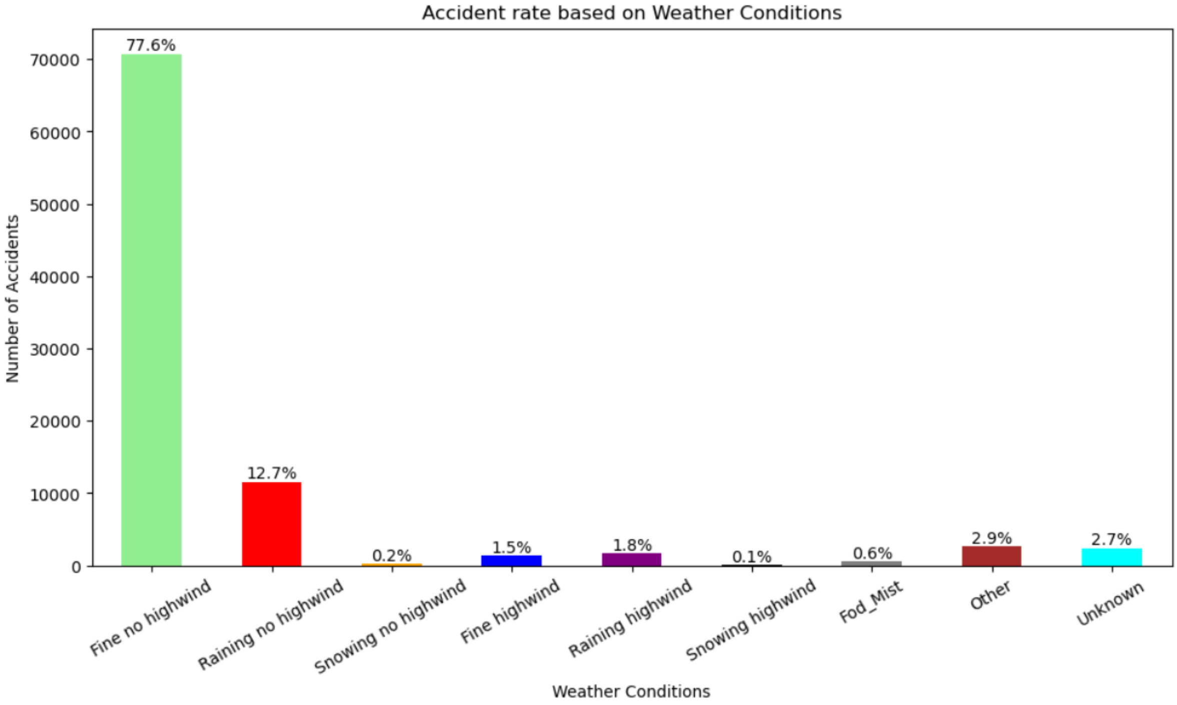


Fig. 14 Effect of weather conditions on accident.

5. Exploring the impact using the apriori algorithm.

Initially, some features were selected for examining their impact on accidents using the apriori algorithm. The selected features were speed limit, weather conditions, and urban or rural areas. The apriori algorithm was applied with minimum support of 0.3, minimum confidence of 0.7, and minimum lift of 1.05. The table 1 represents the best rules that illustrate the relationship between various factors and the slight accident occurrence. For instance, with a support of 35.97%, confidence of 79.91%, and a lift of 1.02, Rule 1 suggests that a speed limit of 30 mph is associated with a higher chance of minor accidents. This indicates that minor injuries are more likely to occur in accidents that occur at slower speeds.

Rule	Rule Body	Support	Confidence	Lift
1	{Speed 30} -> {Slight}	0.359697	0.799084	1.019911
2	{Fine weather no wind, Speed 30} -> {Slight}	0.359697	0.799084	1.019911
3	{Fine weather no wind, Urban} -> {Slight}	0.359697	0.799084	1.019911
4	{Urban, Speed 30} -> {Slight}	0.459983	0.802717	1.024548
5	{Fine weather no wind, Urban, Speed 30} -> {Slight}	0.359697	0.799084	1.019911

Rule 2 implies an association between fine weather conditions with no wind, a speed limit of 30 mph, and slight accidents, emphasising the effect of lower-speed environments on minor accidents. The influence of urban environment on accident severity is portrayed in Rule 3. The minor accidents are more frequent in urban areas, especially those with a 30-mph speed limit. These findings suggest that multiple combinations of speed limits, weather conditions, and urban environments may have an impact on the frequency of minor accidents.

6. Humberside region Clustering.

The distribution of accidents in the Humberside region was examined using K-Means clustering algorithm for clustering the longitude and latitude data. The optimal number of clusters was determined using the elbow curve resulting 5 clusters. From the clusters, it’s evident that most of the accidents occur in nearby areas of major Humberside cities, including Grimsby, Scunthorpe, Bridlington, and Hull (Fig. 15).

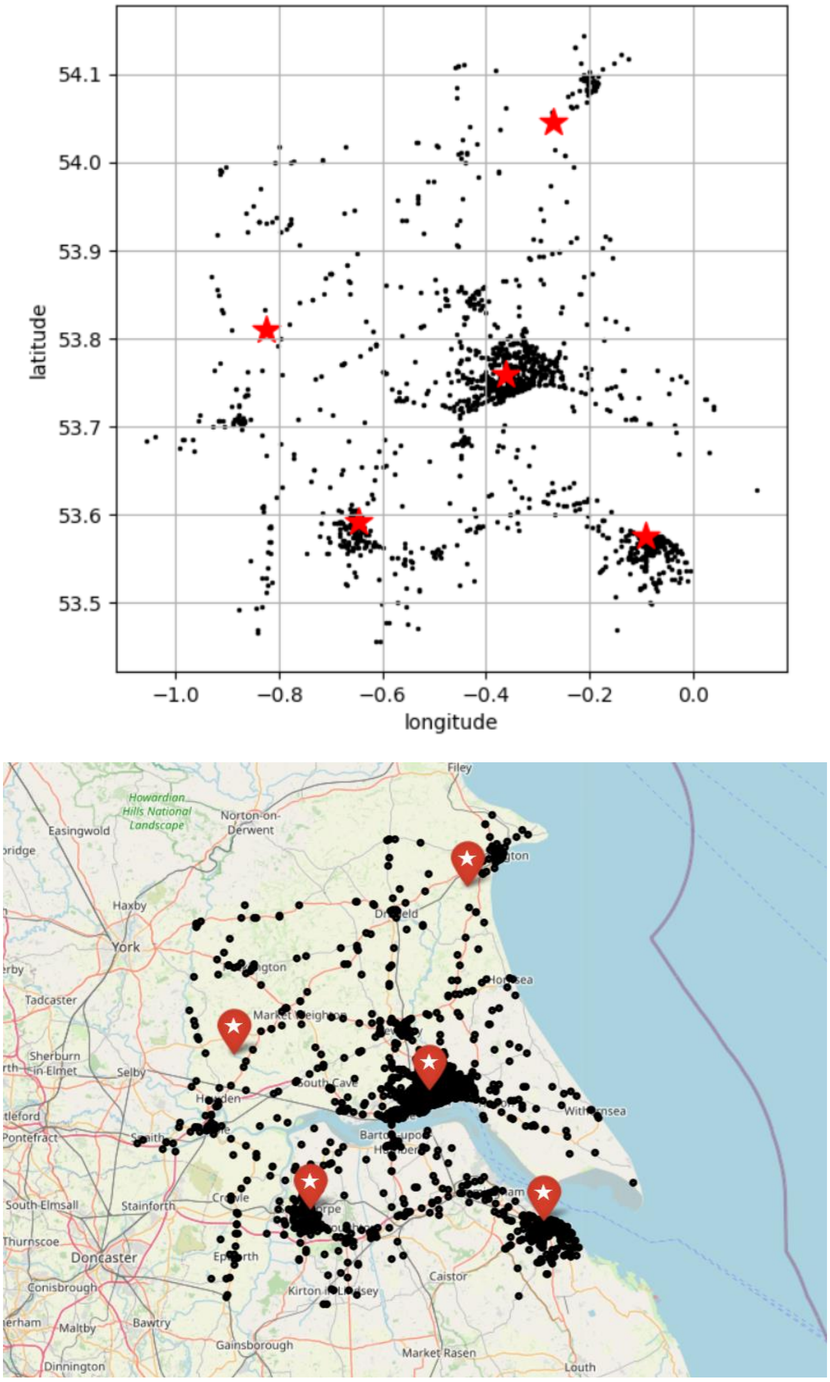


Fig. 15 Clustering corresponds to accident occurrence in Humberside Region

In addition to that, clustering was also carried out based on weather conditions and speed limit. The clusters were associated with weather condition 2 with number of clusters 5, indicating raining without high winds. While increasing the number of clusters from 5 to 10 revealed insights about the impact of fog and other weather-related factors on accidents (Fig. 16).

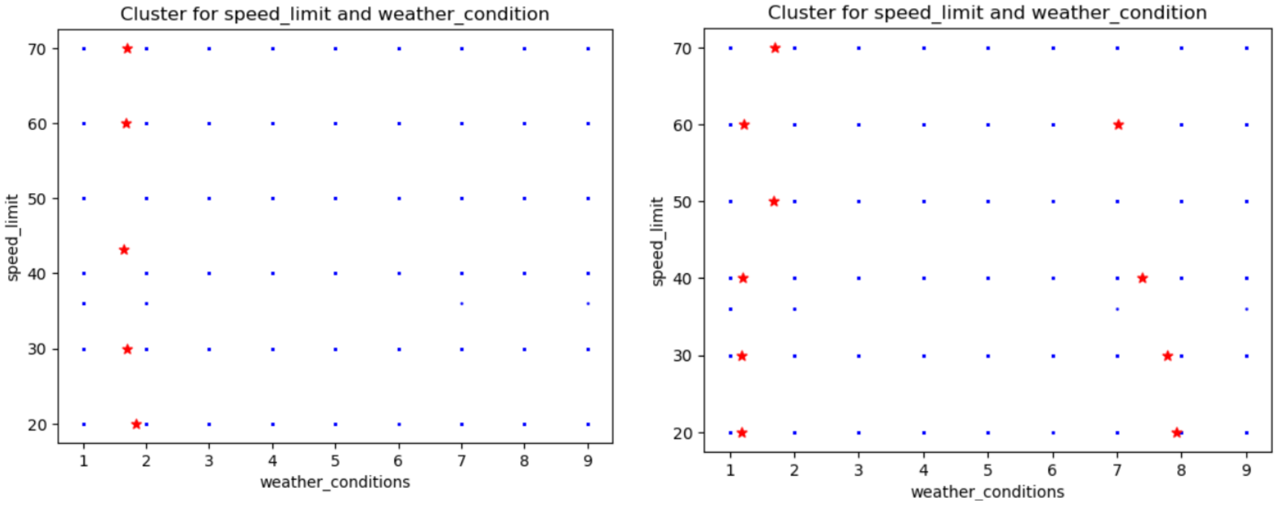


Fig. 16 Clustering based on speed limit and weather condition.

7. Outlier Detection.

The outlier detection was done by utilising the Local Outlier Factor (LOF) and Interquartile Range (IQR) methods. The analysis revealed that outliers have been distributed throughout the dataset, with minor concentration in and near the urban areas. These outliers may be attributed to certain circumstances (Fig. 17).

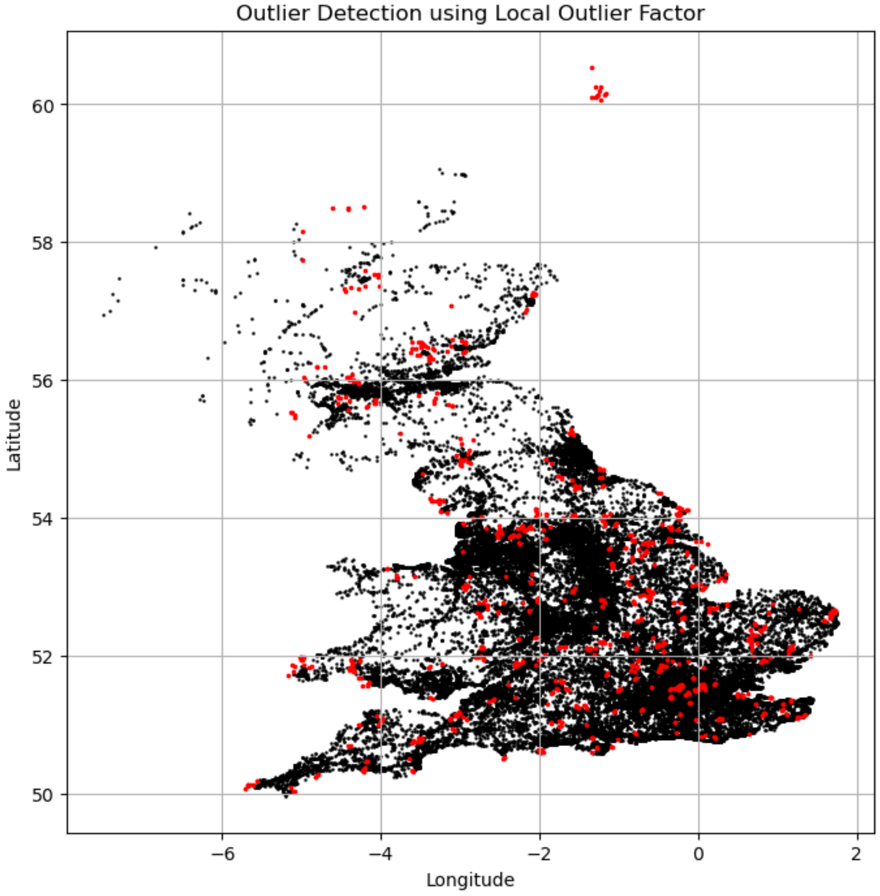


Fig. 17 Outlier Distribution using Local Outlier Factor.

Upon analysing outliers associated with the Humberside region using the IQR method, the results revealed that the outliers are not closely located to the major cities in the Humberside region (Fig. 18). It was decided to retain the outliers in the dataset since these outliers might be correlated with specific circumstances or scenarios.

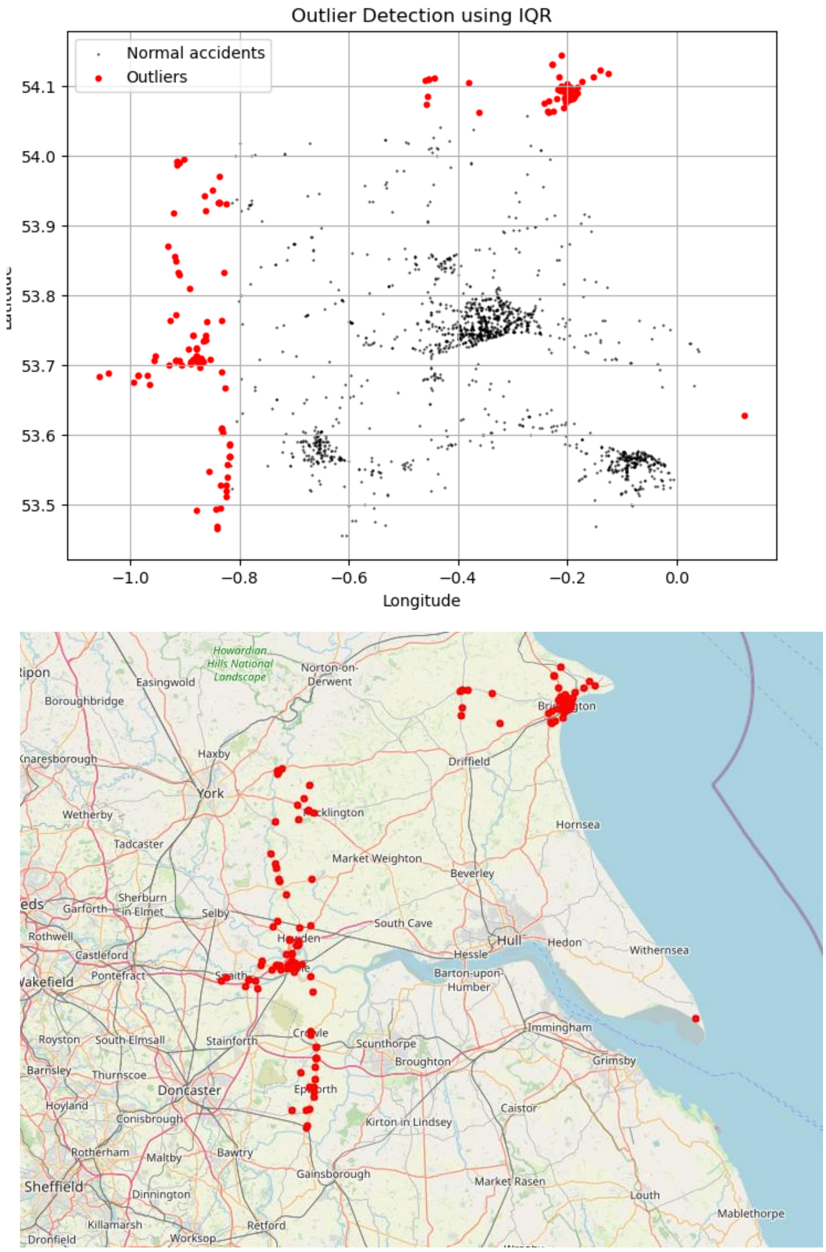


Fig. 18 Outlier Distribution in Humberside Region Using IQR.

Prediction

The most relevant features for the development of the prediction model were selected based on importance indices using the Random Forests. The visual representation of feature importance is shown in Fig. 19. Based on their importance; nine features were chosen. The selected features consist of both numerical and categorical features. Thus, by using LabelEncoder, the categorical features were encoded. Afterwards, the data was split into training and testing sets, followed by standard scaling. A classification model was then constructed with the following hyperparameter, n_estimators=300.

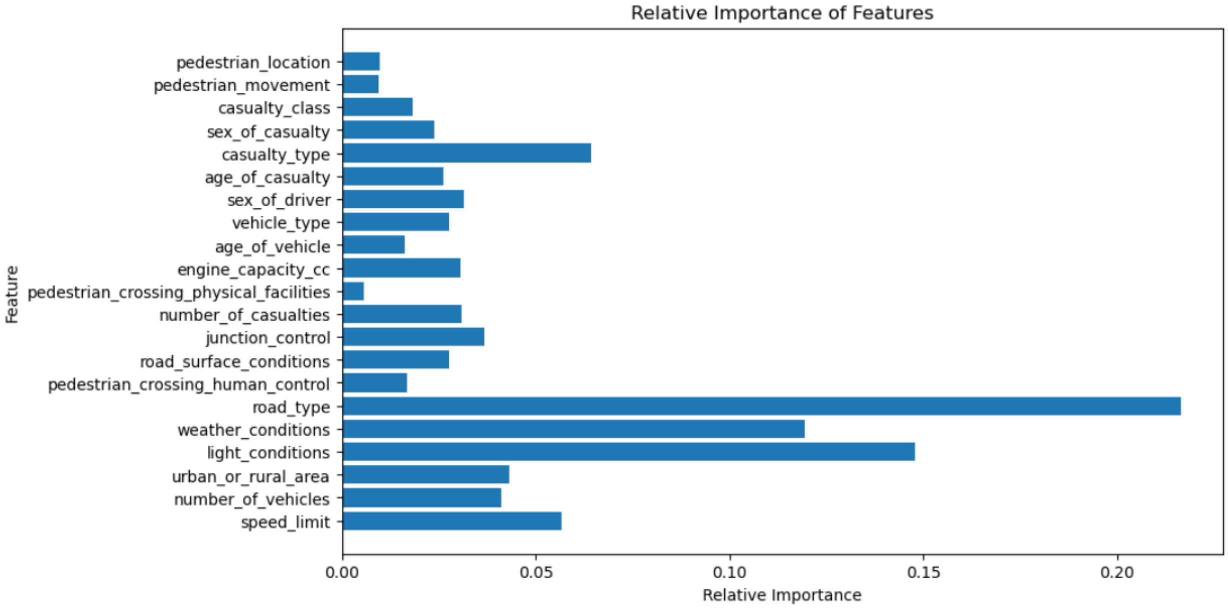
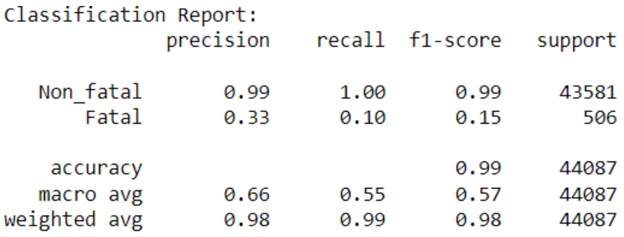


Fig. 19 Relative Feature Importance.

From the Random Forest analysis, the model accurately classified 49 out of 506 fatal accidents as fatal, resulting a low recall and precision of 33%. However, the model achieved a precision of 99% in classifying 43,480 of the 43,581 non-fatal accidents as non-fatal. But, the model achieved an overall accuracy of 99%, indicating its ability to generalise effectively, with varying precision, recall, and F1-score (as shown in Fig. 20).



```
Classification Report:
              precision    recall  f1-score   support

   Non_fatal       0.99      1.00      0.99     43581
       Fatal       0.33      0.10      0.15       506

    accuracy                           0.99     44087
   macro avg       0.66      0.55      0.57     44087
weighted avg       0.98      0.99      0.98     44087
```

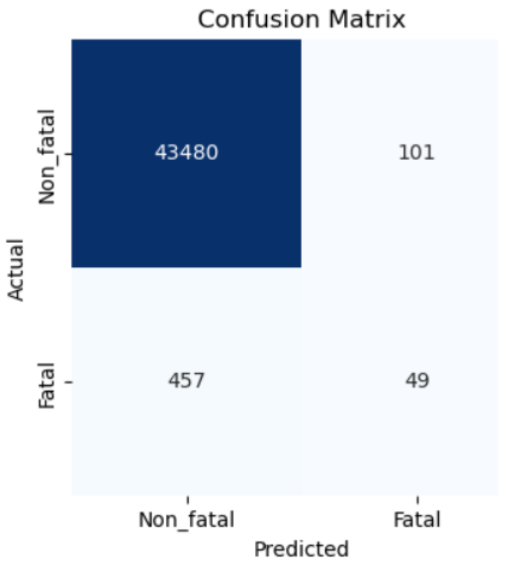


Fig. 20 Classification Report and Confusion Matrix

Recommendations

Based on the data insights, following suggestions are proposed to the government agencies to improve the road safety.

- **Targeted interventions during peak hours**: Significant number of accidents was recorded during the peak hours mainly between 3 PM and 5 PM. Thus, targeted interventions, between these peak hours, such as police presence and any traffic management strategies will reduce the number of accidents.
- **Implementing speed limit enforcement in urban areas**: There exists a need for improving the speed limit enforcement in urban areas. From the association rules, it's clear that the speed limit of 30 mph is associated with minor accidents. Even though it results in minor accidents, the situation must be dealt with the utmost care. This can be achieved by establishing speed cameras, and increased police patrols.
- **Enhancing motorbike safety measures**: The motorbike accidents are concentrated, primarily on Fridays and Sundays, between 2 PM and 6 PM. Thus, motorbike safety measures such as specific motorbike lanes and enhanced road signs should be implemented. Also, an awareness should be made for motorbikes over 500cc and promote the use of motorbikes with moderate engine power.
- **Improving pedestrian safety infrastructure**: There exists a need for more enhanced pedestrian safety infrastructure. Pedestrian-friendly Road signs or designs and illuminated cross walks will reduce pedestrian accidents.

References

Government Digital Service (2015) Driving lessons and learning to drive. https://www.gov.uk/driving-lessons-learning-to-drive#:~:text=You%20can%20apply%20for%20a,Personal%20Independence%20Payment%20(PIP).