Introduction

Accidents happen due to numerous factors, which can serve as attributes to comprehend the patterns of accidents and predict their likelihood. Similarly, data from road traffic accident reports provide insights into the circumstances surrounding accidents and the human consequences of each incident. Therefore, these records offer insights to suggest ways to minimize risks and promote responsible behavior when needed. The primary aim of this project is to provide data-driven suggestions for enhancing road safety. It involves building models that can anticipate accident probabilities and the associated injuries, contributing to a more informed approach toward accident prevention and mitigation.

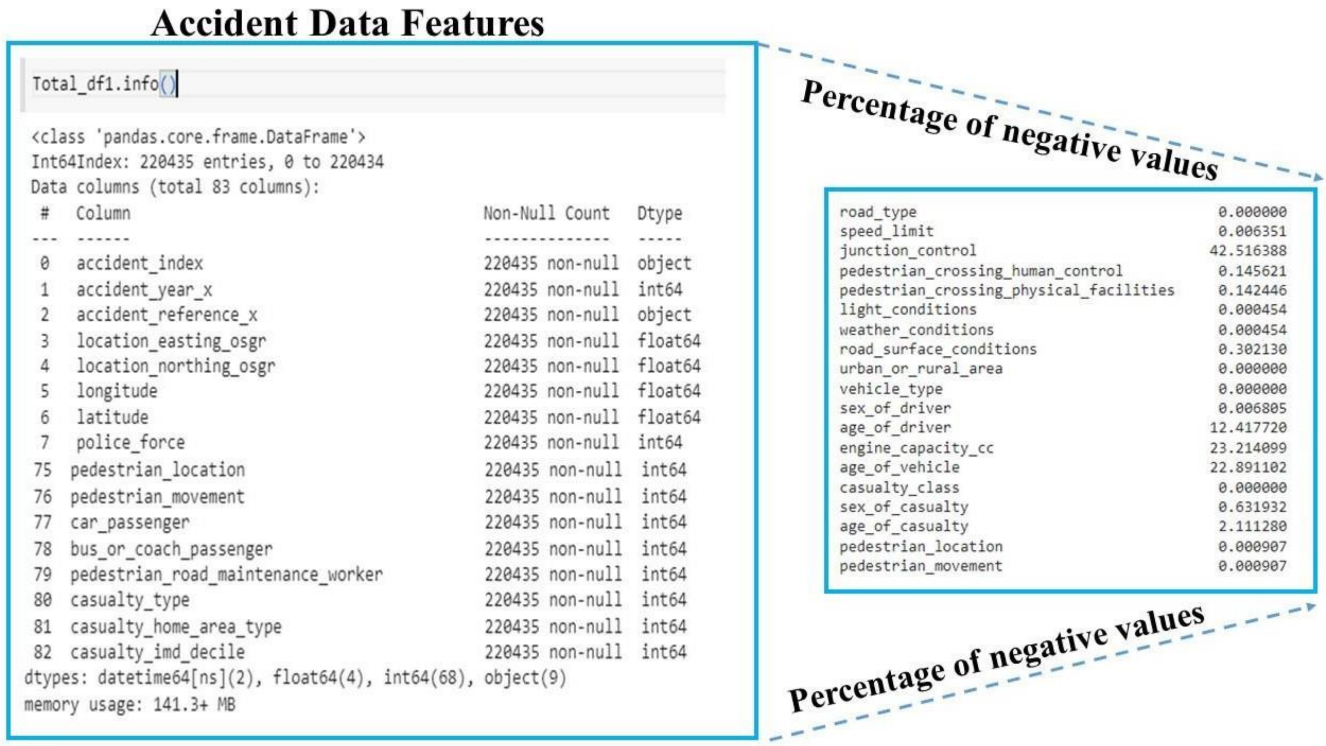
Methodology

Data preprocessing (exploratory data analysis and cleaning) was first conducted to enhance prediction accuracy. Afterward, the Apriori algorithm was utilized for association mining and clustering to comprehend accident causes better. The data was then normalized for model construction before proceeding with feature selection. Ultimately, the Random Forest, decision tree, and Gradient Boost algorithms were used to predict traffic accident severity.

Data Description and cleaning

The study employed the UK accident database from 2020, consisting of four tables: accident, table, vehicle, casualty, and LSOA of the accident location. Upon merging these tables, the dataset contained 82 columns and 220,435 rows (as depicted in Fig. 1). Exploratory data analysis using Sweetviz revealed 14 NAN values across columns like "location_easting_osgr," "location_northing_osgr," "longitude," and "latitude." Notably, some missing entries were labeled as -1. Fig. 1 illustrates the proportions of negative values in some important features. Data cleaning focused on features that were pivotal to our project. The missing latitude and longitude values were found to correspond to distinct police stations, as accidents are managed by separate police units within specific regions. These voids were filled using location data from the corresponding police stations.

The average replacement method was applied for numeric features with negative values (e.g., vehicle age, engine capacity, casualty age). Categorical columns with negative values were substituted with the mode. Special consideration was taken to clean the age of the driver. For instance, the Age of driver spanned from 1 to 101, averaging 36. This distribution seemed unreasonable, given that driving in the UK is legal for those aged 17 and above (Gov. UK, 2023). To address this, the median age between 17 and the oldest legal age was used to replace negative values. Additionally, the date column was converted to DateTime format to facilitate analysis. Moreover, new-time, converted_time columns were introduced to the dataset by transforming the time column into hours and minutes.



```
Total_df1.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 220435 entries, 0 to 220434
Data columns (total 83 columns):
 #   Column                              Non-Null Count   Dtype
---  ------                              --------------   -----
 0   accident_index                      220435 non-null  object
 1   accident_year_x                     220435 non-null  int64
 2   accident_reference_x                220435 non-null  object
 3   location_easting_osgr               220435 non-null  float64
 4   location_northing_osgr              220435 non-null  float64
 5   longitude                           220435 non-null  float64
 6   latitude                            220435 non-null  float64
 7   police_force                        220435 non-null  int64
 75  pedestrian_location                 220435 non-null  int64
 76  pedestrian_movement                 220435 non-null  int64
 77  car_passenger                       220435 non-null  int64
 78  bus_or_coach_passenger              220435 non-null  int64
 79  pedestrian_road_maintenance_worker  220435 non-null  int64
 80  casualty_type                       220435 non-null  int64
 81  casualty_home_area_type             220435 non-null  int64
 82  casualty_imd_decile                 220435 non-null  int64
dtypes: datetime64[ns](2), float64(4), int64(68), object(9)
memory usage: 141.3+ MB
```

```
road_type                                     0.000000
speed_limit                                   0.006351
junction_control                             42.516388
pedestrian_crossing_human_control             0.145621
pedestrian_crossing_physical_facilities       0.142446
light_conditions                              0.000454
weather_conditions                            0.000454
road_surface_conditions                       0.302130
urban_or_rural_area                           0.000000
vehicle_type                                  0.000000
sex_of_driver                                 0.006805
age_of_driver                                12.417720
engine_capacity_cc                           23.214099
age_of_vehicle                               22.891102
casualty_class                                0.000000
sex_of_casualty                               0.631932
age_of_casualty                               2.111280
pedestrian_location                           0.000907
pedestrian_movement                           0.000907
```

Fig.1 The accident data

Accident Demography

Considering the geographical aspect of traffic accidents, Folium was utilized to depict the occurrence of high-density accident hotspots across the UK. The density plot shows a significant concentration of accidents within major urban cities in England, notably London, Manchester, Liverpool, Leeds, and Birmingham. These cities also correspond to the most populous areas in the UK (Feng et al., 2020). Furthermore, the analysis reveals that more than 75% of accidents occur on single carriage roads within urban cities, as exemplified by the clustering of red dots on the map in inner city areas (Fig. 2). It could be the reason for higher local traffic congestion compared to highways. Approximately 78% of recorded accidents are categorized as minor, with fatal injuries accounting for less than 2% of the total accidents.

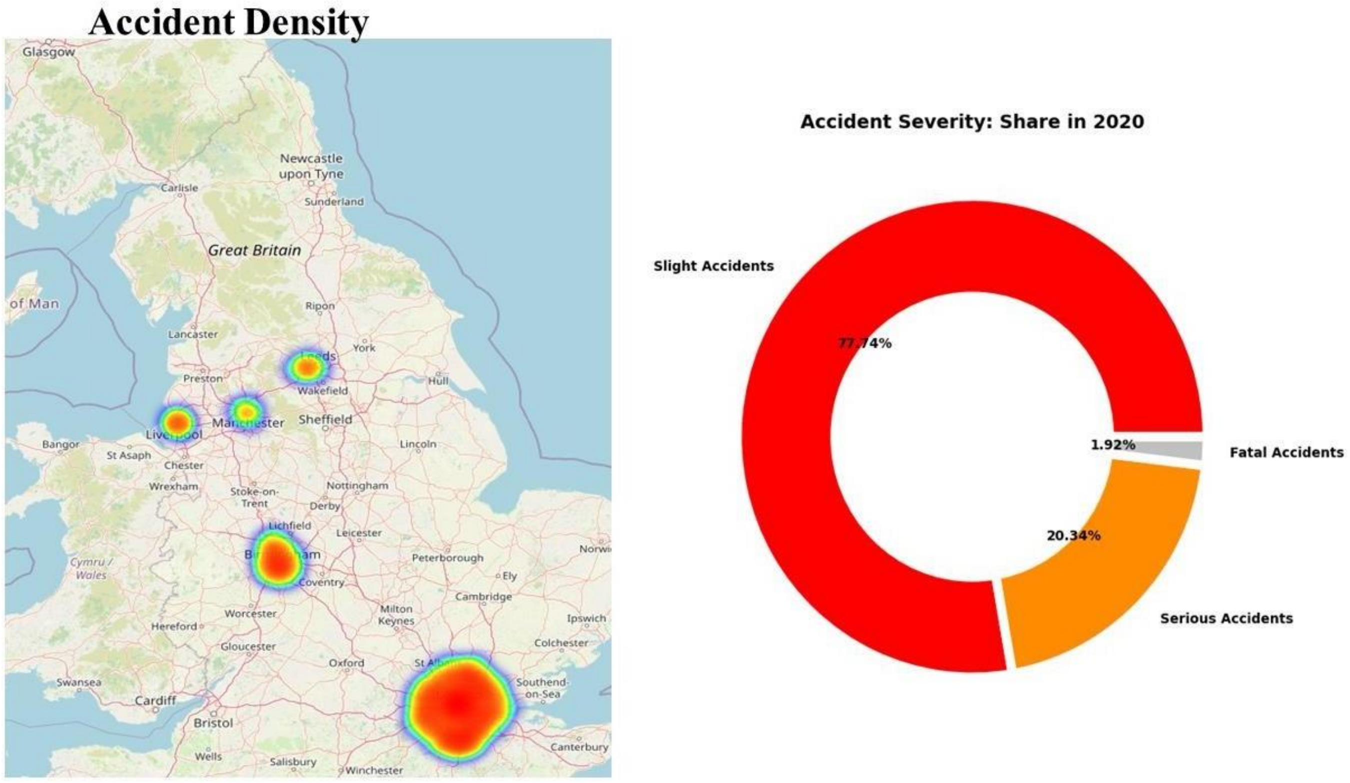


Fig 2 UK accident Hotspot

In general, there were more male casualties and male drivers compared to females. The age group most susceptible to accidents was between 26 and 35 years. Interestingly, most drivers fell within the age range of 21 to 55, and coincidentally, most casualties also occurred within this age bracket, as shown in Fig. 3. This suggests that individuals within this age range are more prone to driving and being involved in accidents.

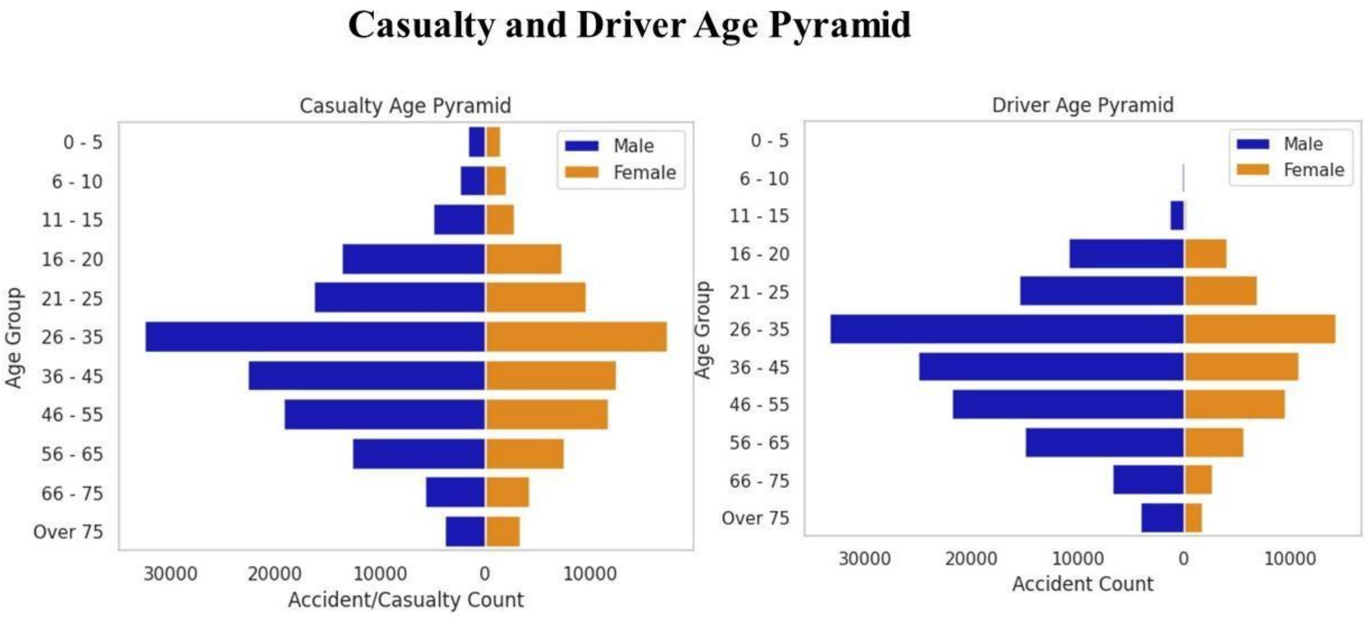


Fig.3 Casualty and driver age pyramid

The time series plot shows the lowest number of casualties were observed during spring (April and May). In contrast, the summer, winter, and Christmas holidays exhibit a substantially higher number of casualties (Fig 4). This trend can be attributed to increased human mobility, particularly during the summer and winter vacations.

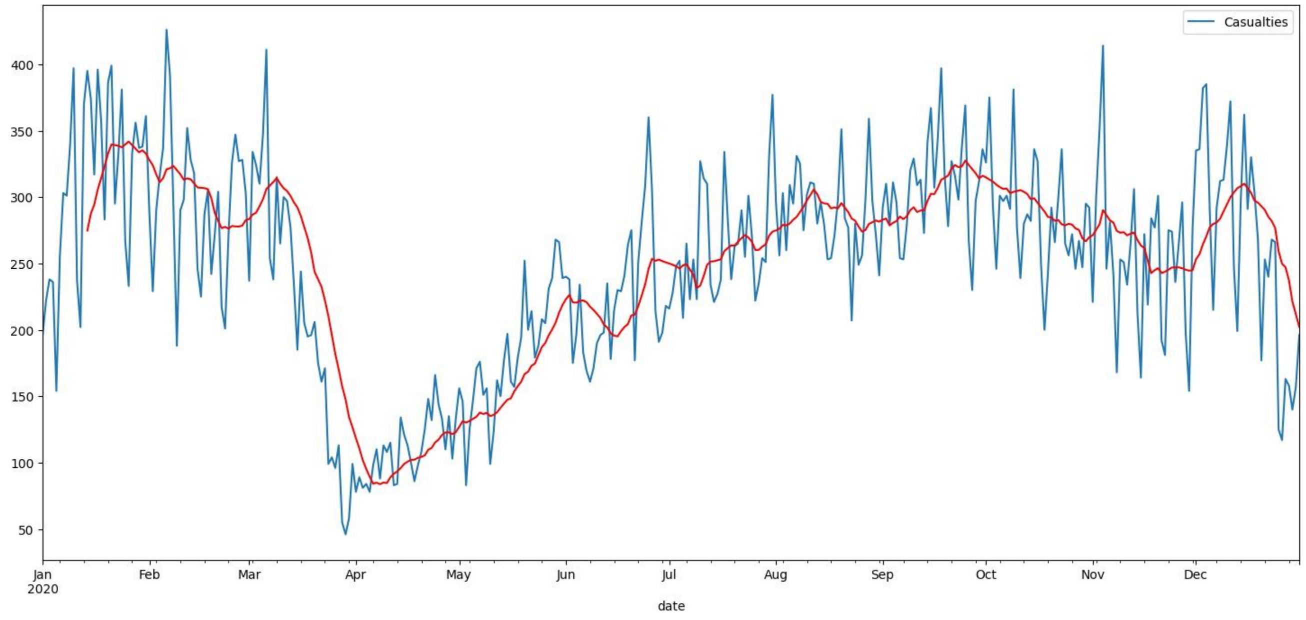


Fig.4. Accident casualty variation during the year.

Data insight

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

Many accidents are observed in the morning, particularly around 8 am. However, a disproportionately high incidence of accidents occurs between 3 to 4 pm, reaching its peak at approximately 5 pm. This temporal trend aligns with rush hour, coinciding with morning and evening commutes as people travel to and from work (see Fig. 5).

Most accidents occur during the weekdays, accounting for approximately 75% of the total accidents, with Friday having the most accident occurrence in the week (about 16%). This spike corresponds to when individuals rush home for the weekend, leading to increased road vehicular activity (Fig. 6).

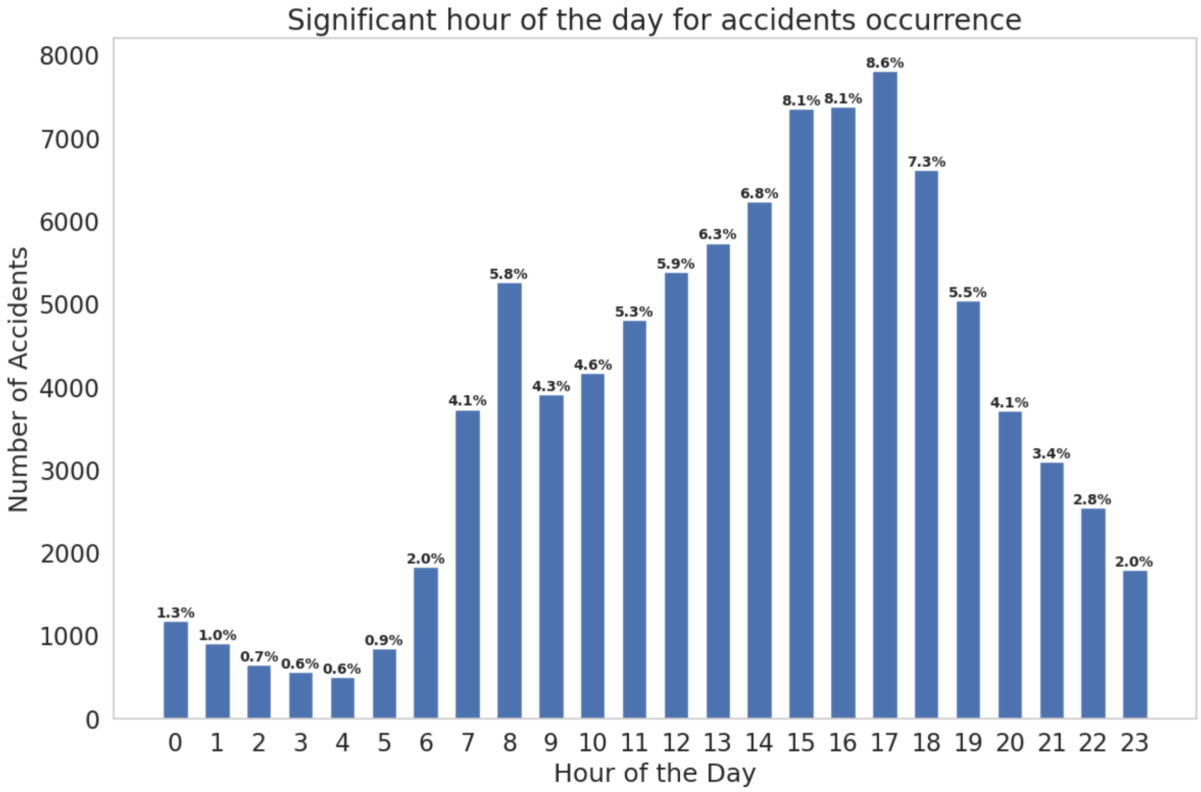


Fig 5 Accident occurrence during the hours of a day

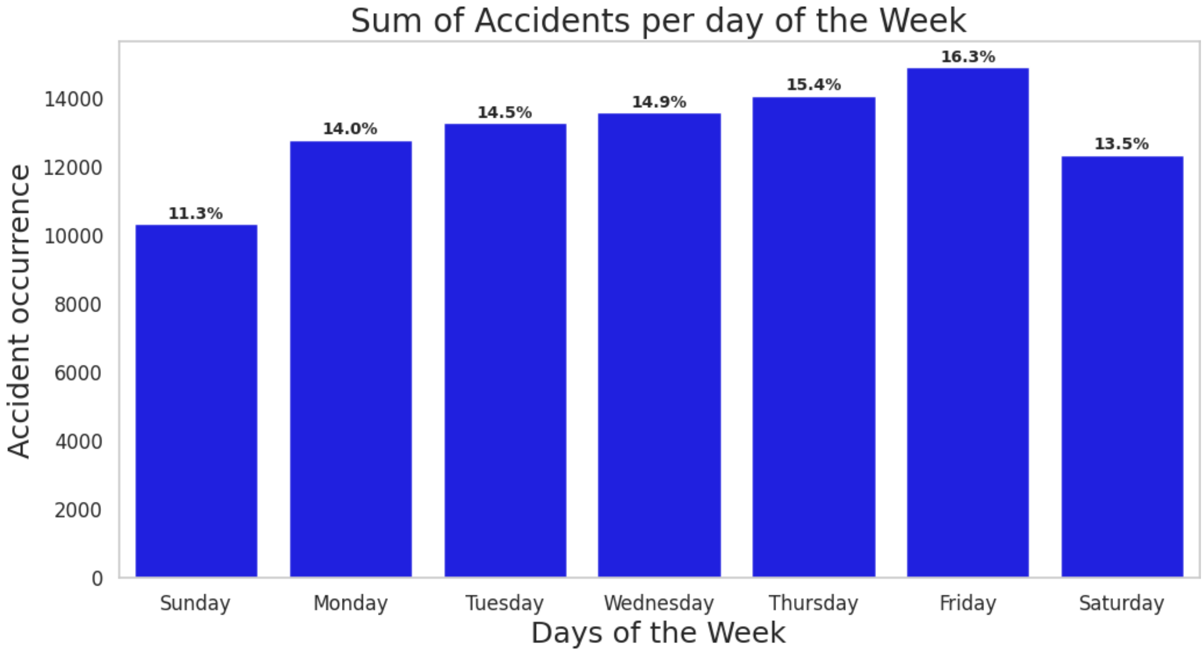


Fig. 6. Accident occurrence per weekday

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

Likewise, motorcycle accidents are primarily concentrated between 3 pm and 6 pm, with a prominent peak observed around 5 pm, as shown in Fig 7. Among various motorcycle types (50cc and under, 50cc to 125 cc, 125cc to 500cc, and over 500cc), the group with 50cc to 125 cc motorcycles records the highest accident involvement. Across all motorcycle categories, Fridays stand out as the day with the highest accident occurrences, except for the group with motorcycles over 500cc, which experiences its peak on Sundays (Fig.8).

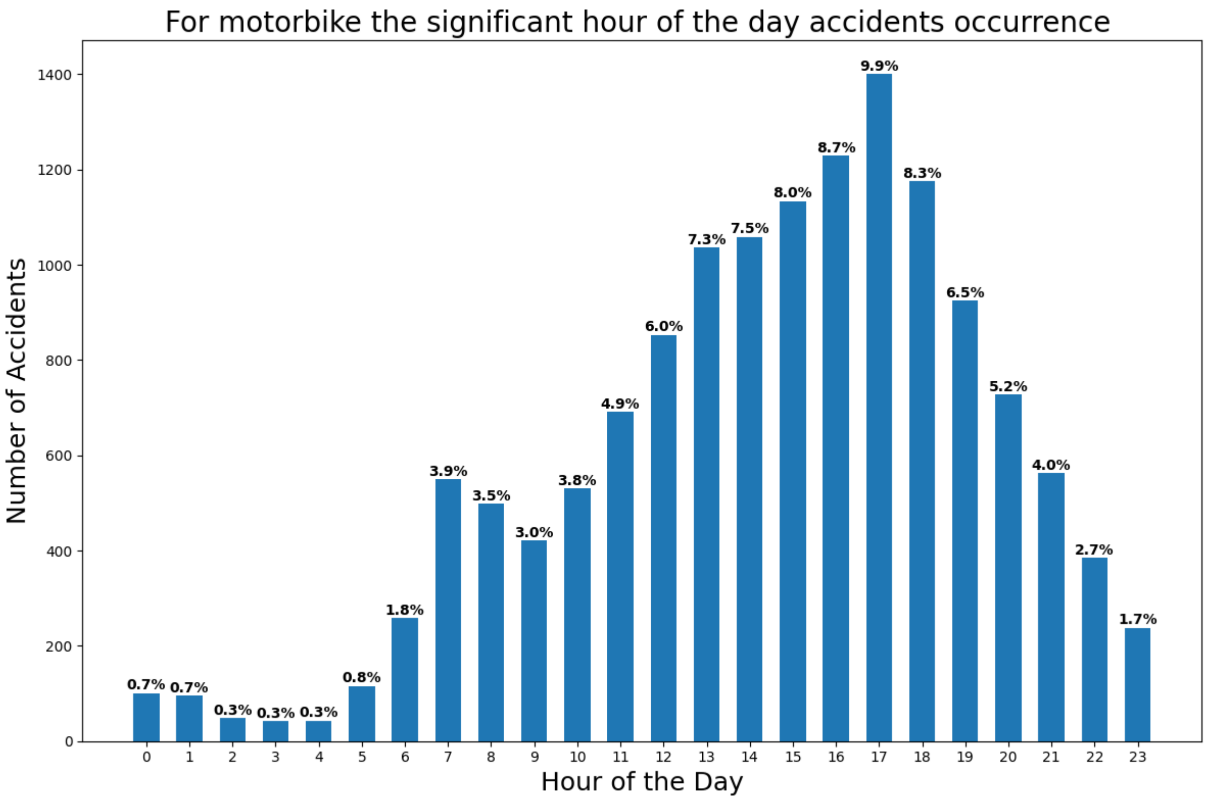


Fig.7 For motorbikes, the significant hour for accident occurrence

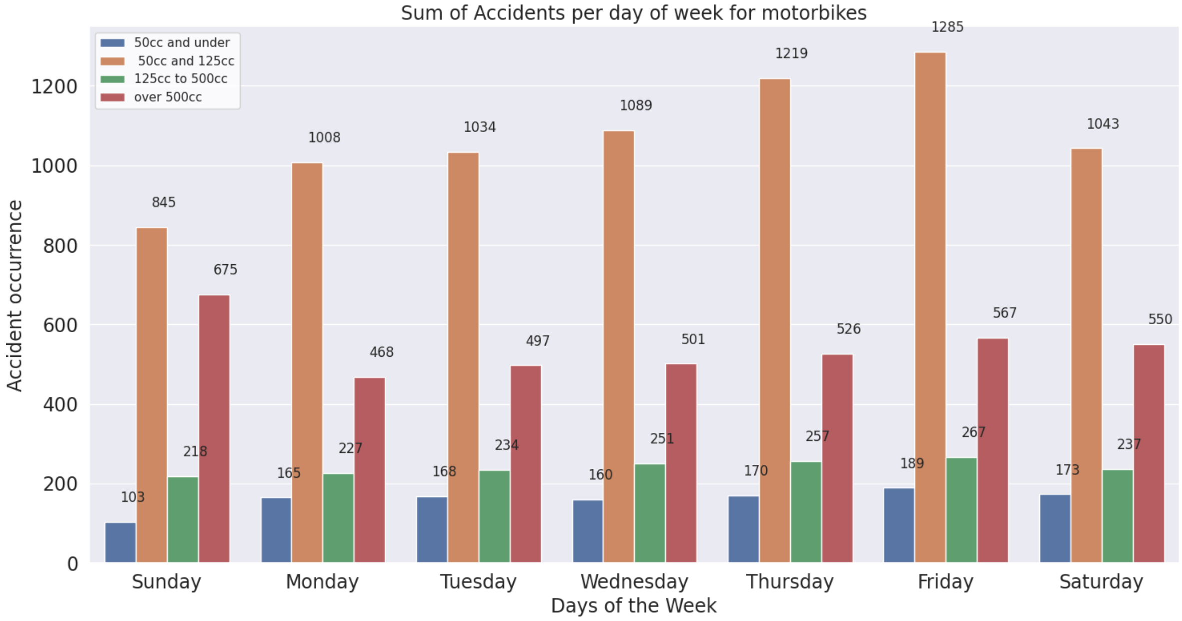


Fig.8 Accident occurrence per weekday for motorbikes

3. The significant hours and days of the week when pedestrian accidents occur.

Most pedestrian accidents occurred during weekdays (around 80%), with an even distribution within the weekday (about 15.5%), and only a slight peak of 17% observed on Fridays (Fig.9). A substantial proportion of these accidents also took place between 3 and 6 pm, peaking at approximately 3 pm (Fig.10.). Additionally, significant accident occurrences were noted around 8 am during the early rush hour period. Among the days of the week, Sundays experienced the least pedestrian accidents. Comparatively, the number of accidents involving drivers or riders was significantly higher, often double or triple that of pedestrian accidents on most days.

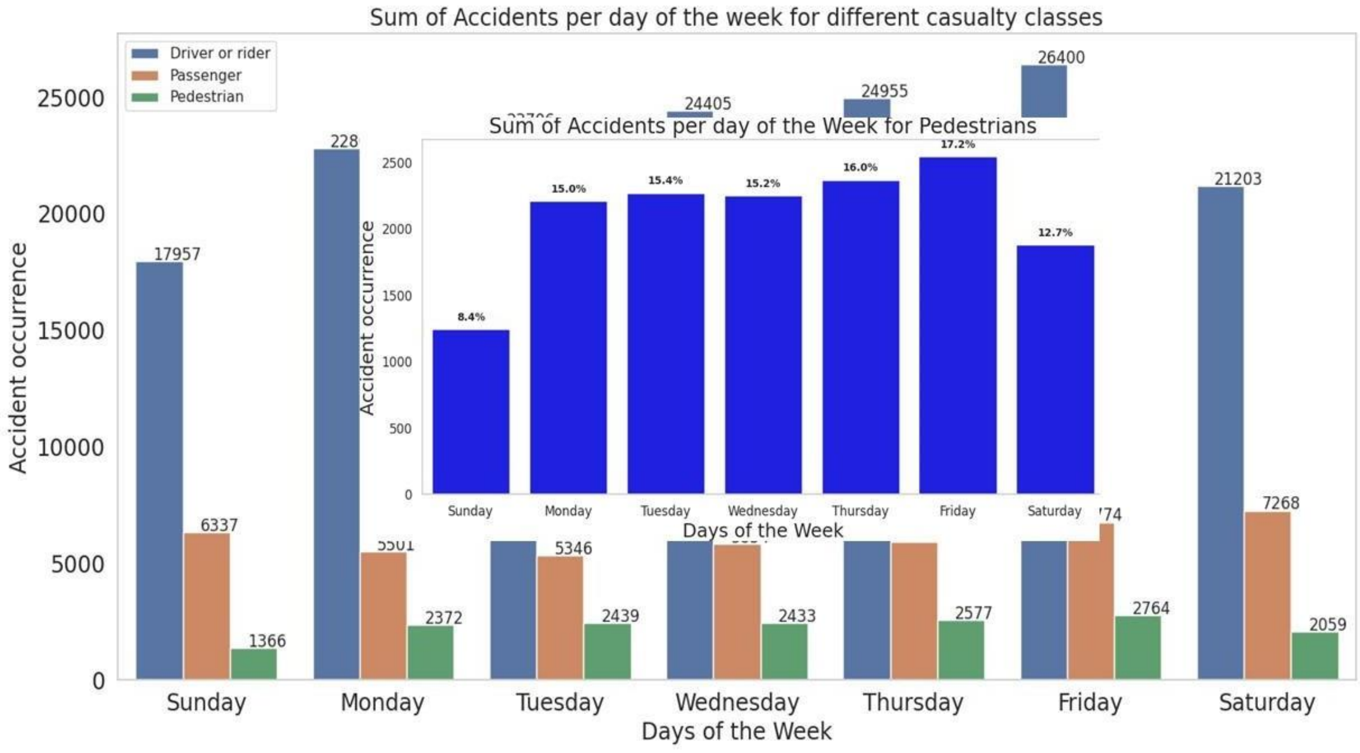


Fig. 9 Distribution of accident occurrence per weekdays for different casualty types

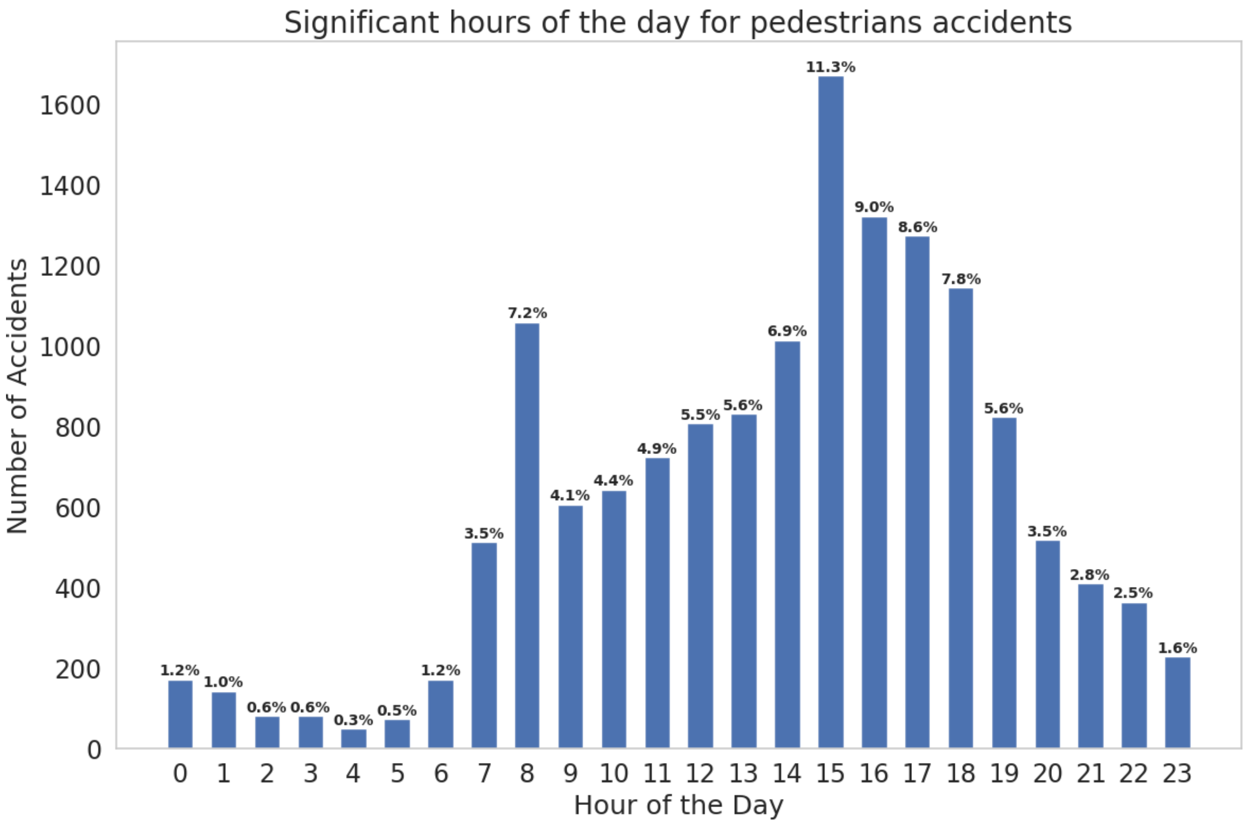


Fig.10 Significant hours of the day for accidents involving pedestrians.

4. Effect of light and weather condition

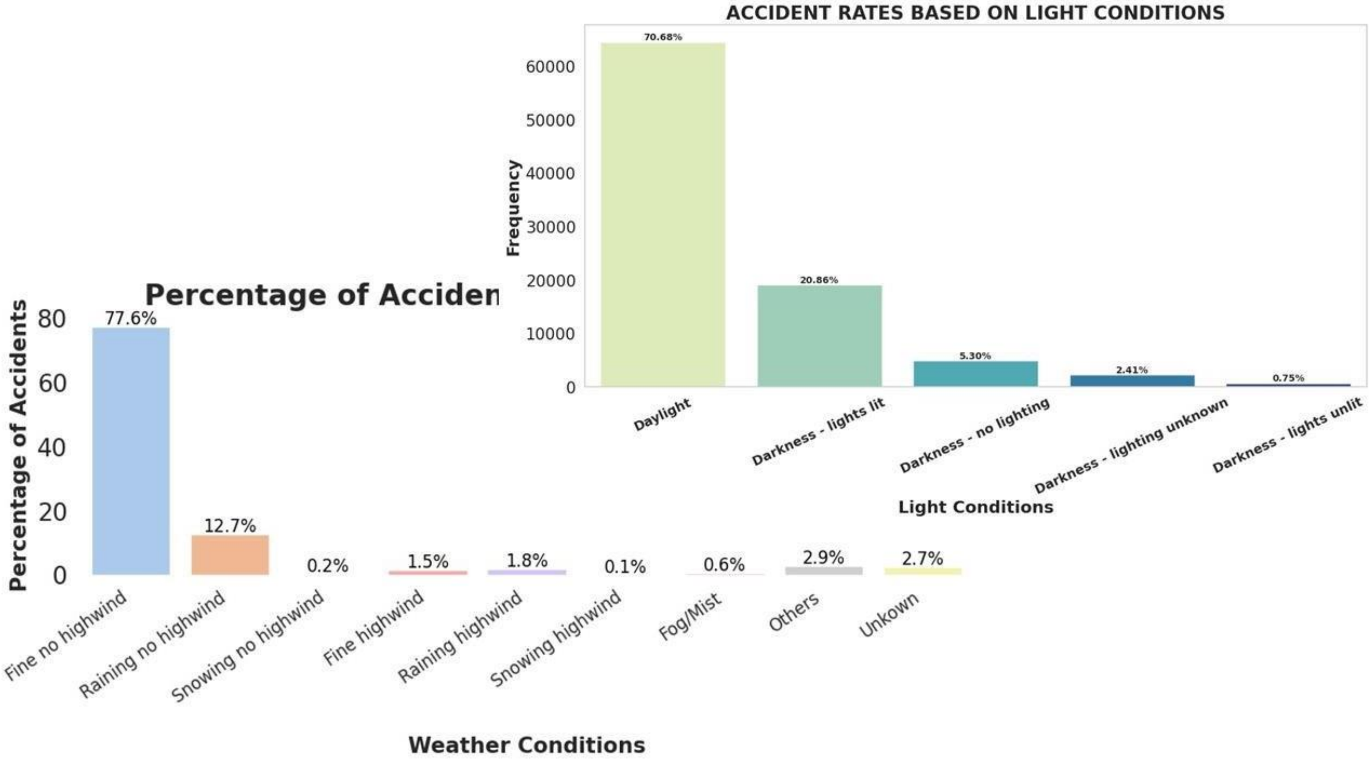


Fig.11 Effect of light and weather conditions on accident

More than 70% of accidents occurred during daylight and under fine weather conditions. Surprisingly, adverse weather conditions like harsh weather and darkness had minimal impact on accident occurrence, contributing to less than 7% of all accidents Fig11. Rather than causing more accidents, the poor weather and darkness reduce them by discouraging driving, leading to fewer drivers and less congested roads, potentially reducing the overall number of accidents.

5. Exploring the impact of selected variables on accident severity using the apriori algorithm

After applying the Apriori algorithm with minimum support of 0.27 and minimum confidence of 0.7, association rules with accident severity categorized as "slight" were generated. The best rules are shown in Table 1. Rule 1 indicates that Urban areas and daylight conditions correlate with slight accidents (37.07% support), (82.78% confidence), and a Lift of 1.065. The result implies that most accidents that occur in urban areas during daylight result in minor injuries.

Table 1 presents five associations with high confidence and lifts discovered by the Apriori algorithm.

Rule	Rule body	Support	Confidence	Lift
1	{Urban, Daylight}→{Slight}	0.370658	0.827788	1.064755
2	{Urban, Daylight, Speed_30}→{Slight}	0.268914	0.825817	1.062220
3	{Urban, Daylight, Dry road}→{Slight}	0.287835	0.825310	1.061567
4	{Urban, Daylight, Dry road, Find weather no wind} →{Slight}	0.270021	0.823367	1.059068
5	{Urban, Daylight, Fine weather no rain} {Slight→}	0.303572	0.822857	1.058412

Rule 2 indicates that Urban, daylight conditions with a speed limit of 30 km/h also relate to slight accidents. Implying lower-speed urban areas influence minor accidents. Rule 4 implies that Urban areas with daylight, dry roads, and fine weather associate well with slight accidents (27.00% support, 82.34% confidence). Implying multiple favourable conditions for the occurrence of slight accidents.

6. Clustering Humberside region

After experimenting with various clustering algorithms (DBSCAN, K-medoids), the K-means algorithm with Euclidean distance was utilized for clustering longitude and latitude to analyze accident distribution (Li et al., 2017). Cluster count was determined using the elbow method, resulting in 5 clusters. Predominantly, accidents clustered around major cities in the Humber region, like Hull, Scunthorpe, and Bridlington (Fig. 12). Clustering speed limits and weather conditions highlighted most clusters under weather condition 2 (rainy), while increasing clusters to 15 revealed the effect of fog on accidents (Fig. 13).

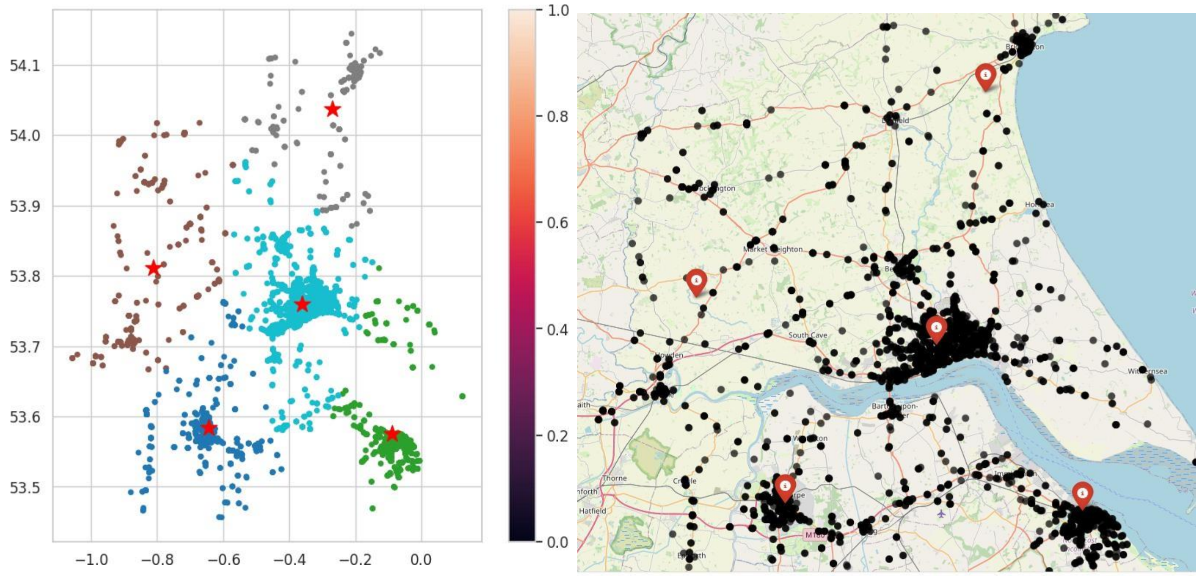


Fig 12**.** The cluster of accident occurrence in Humberside region

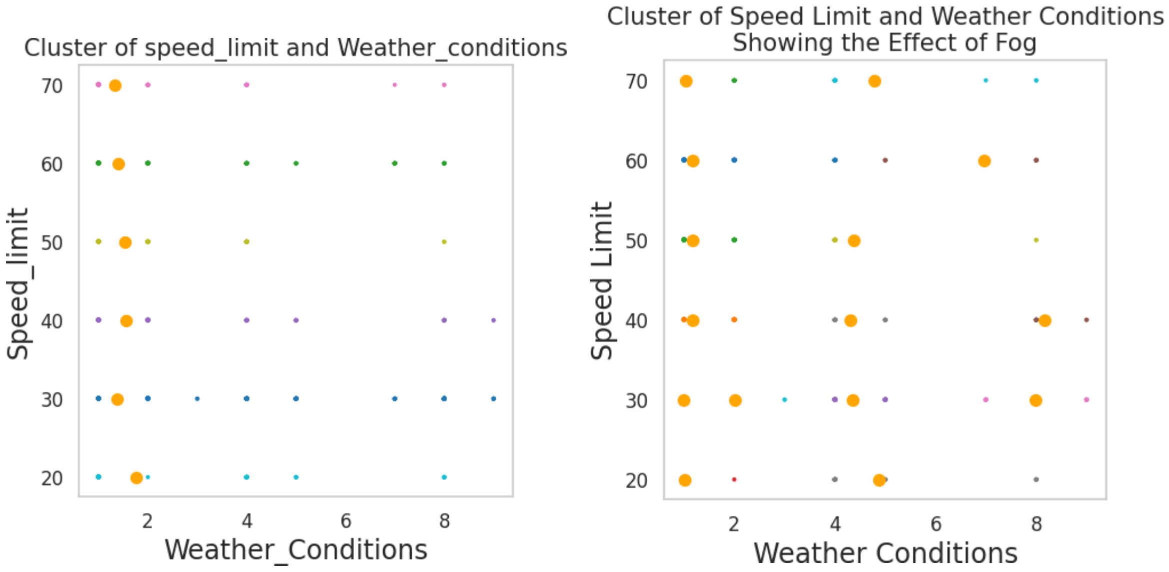


Fig.13 Clusters showing the effect of speed limit and weather conditions on accident occurrence.

7. Outlier Detection.

Outlier detection was performed on specific accident data features using IQR, Isolation Forest, and Local Outlier Factor techniques. Geospatial analysis revealed that outliers were distributed across the dataset, with a minor concentration near the London area Fig 14. The localized clustering of these outliers suggests that certain factors or circumstances might contribute to these anomalous incidents. Subsetting the Humberside region (Fig.15) shows the outliers are not extreme values, which means they are likely not caused by data entry errors or outliers arising from measurement inaccuracies. Instead, they might represent unique or uncommon situations that occurred within the city (they are true outliers), so I decided to keep the outliers.

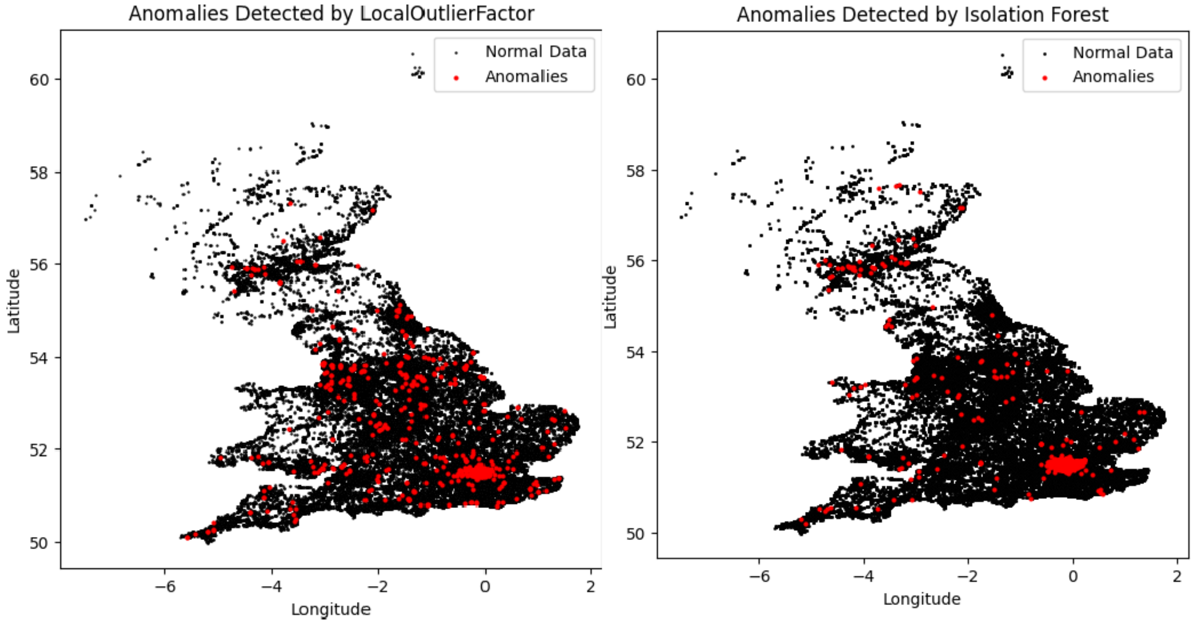


Fig 14 Outlier distribution

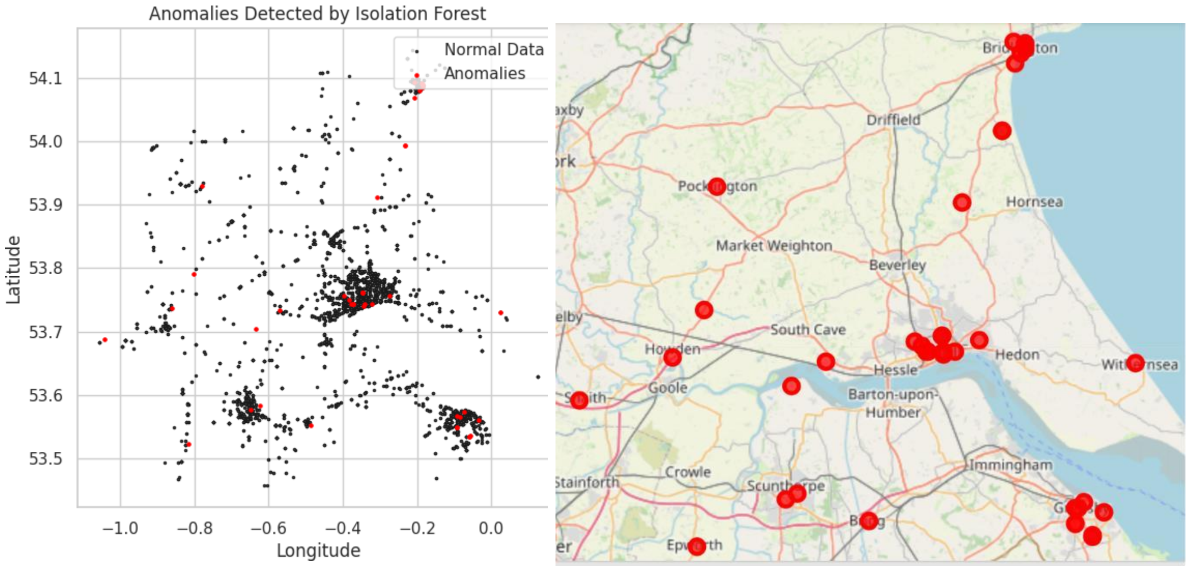


Fig 15 Humberside Outliers

Prediction

To ensure the development of an accurate predictive model, it is essential to apply techniques for feature balancing and normalization (under-sampling and standard scaling) (Haynes et al., 2019). The relevant features were selected using the Random Forests (RFs) based on their importance indices. The visual representation of feature importance ranking from the RFs is depicted in Fig 16. From this selection, the top 10 features were chosen, and an extensive Grid search was executed to identify the most suitable hyperparameters for constructing the classification model(criterion='gini', max_depth=15, min_samples_split=2, n_estimators=300).

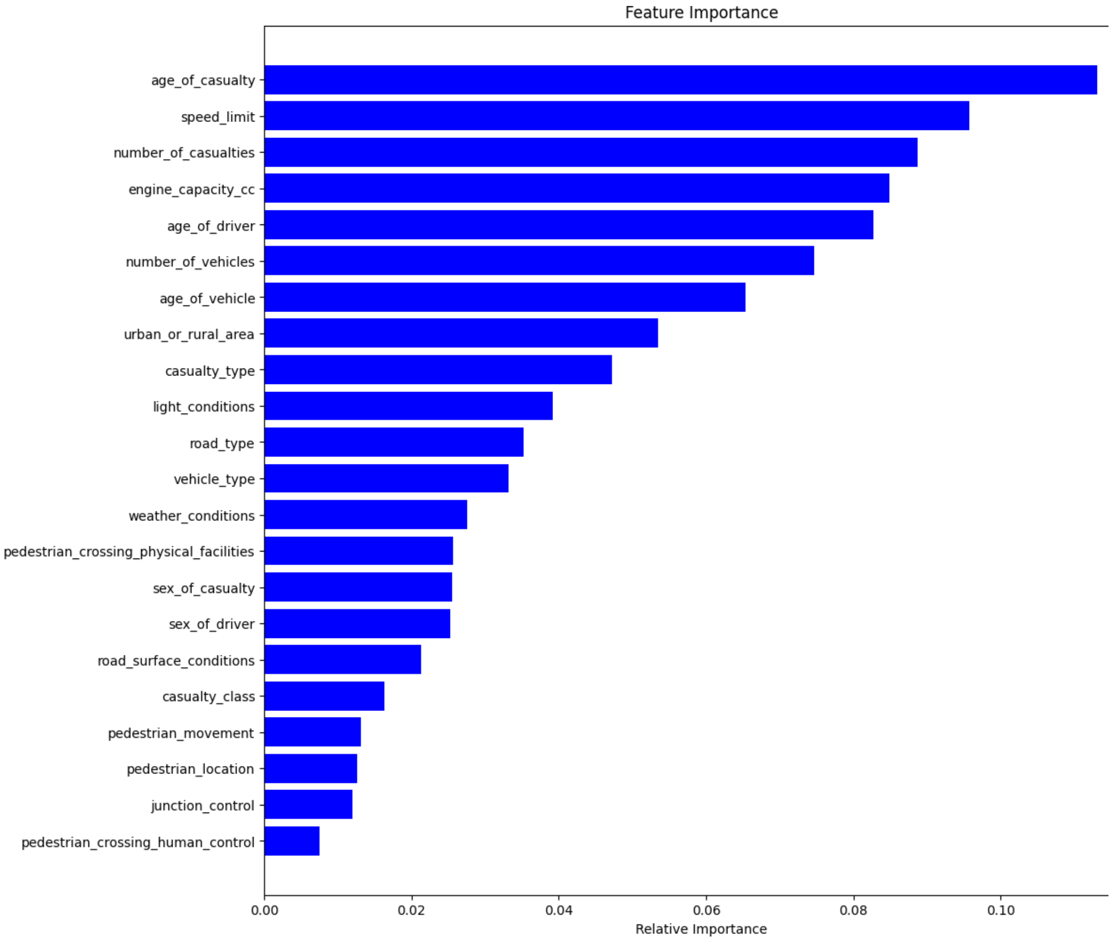
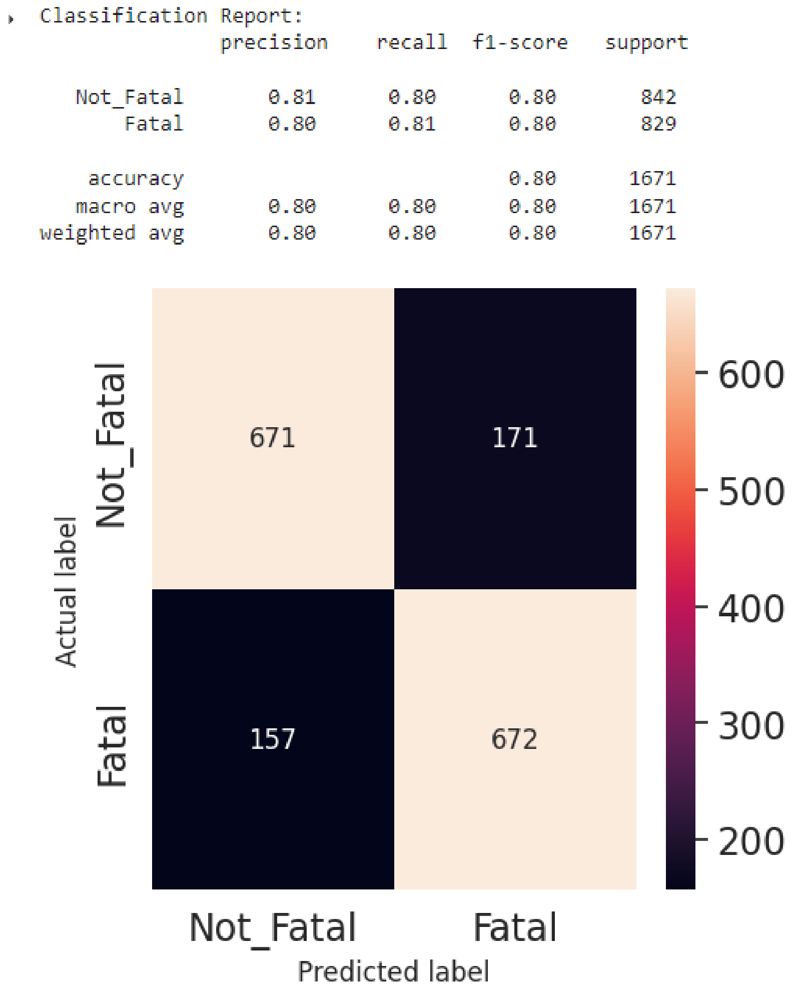


Fig. 16 Feature importance

The outcomes of the Random Forest analysis are presented in Fig. 17. Among the 829 fatal accidents within the test dataset, the Random Forest model achieved a commendable recall rate of 81%, accurately identifying 672 cases. Overall, the model generalized well with accuracy, precision, and F1 score, each at 80% or higher. Cross-validation improved the recall, precision, and F1 score of the random forest model; however, the model accuracy was reduced (table 2).

In a comparative analysis against Gradient Boost and Decision Tree models, the Random Forest model demonstrated superior performance in predicting fatal accidents (Fig .18). Even with stacking, the random forest outperformed every other classification model Fig 19. In summary, the model displayed robust predictive capabilities for identifying fatal accident occurrences with significant accuracy.



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

   Not_Fatal       0.81      0.80      0.80       842
       Fatal       0.80      0.81      0.80       829

    accuracy                           0.80      1671
   macro avg       0.80      0.80      0.80      1671
weighted avg       0.80      0.80      0.80      1671
```

Fig.17 Random Forest classification report and confusion matric

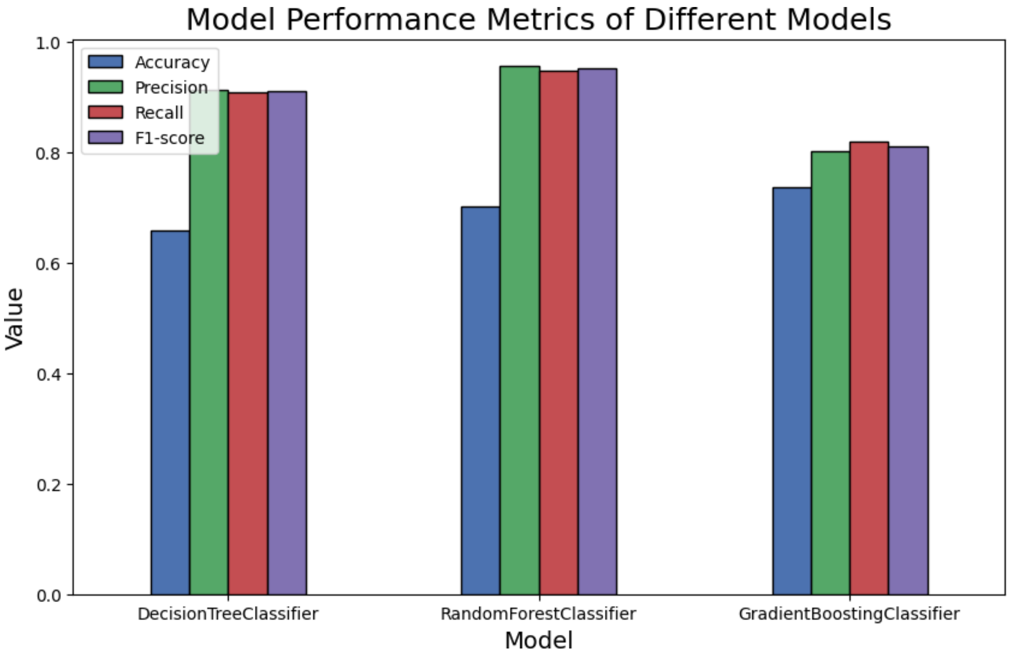


Fig.18 Model performance metric of different classification algorithm

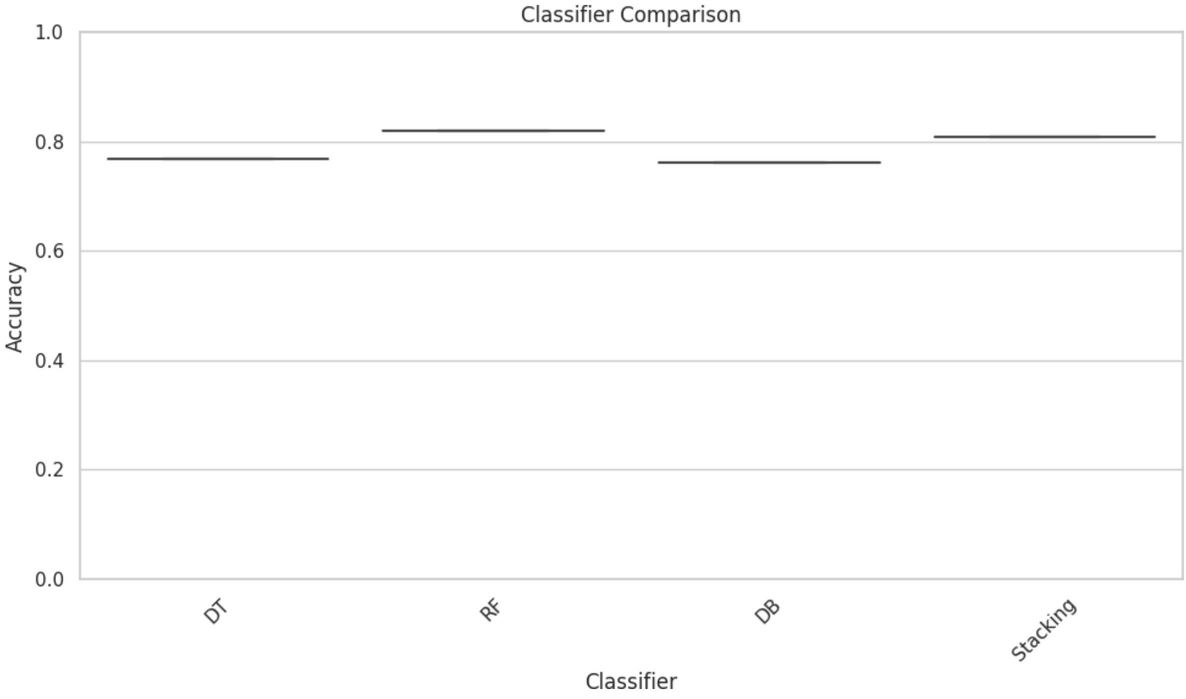


Fig 19 Model performance accuracy for different classifiers.

Table.2. Performance of random forest model with and without cross-validation

RF Model	Accuracy	Precision	Recall	F1-score
Without cross-validation	0.80	0.80	0.81	0.80
With Cross-validation	0.70	0.96	0.95	0.95

Recommendation

Based on the comprehensive data analysis and prediction, the following recommendations are derived:

1. Optimize Urban Roads: Urban areas witness most accidents on single-carriageways. Consider expanding busy roads during rush hours or converting to dual carriageways for safer and quicker travel.

2. Promote Safer Motorcycles: Introduce policies replacing 50cc-125cc and over 500cc motorcycles with safer electric ones. This reduces accidents and carbon footprint.

3. Enforce Scooter Rules: Enact strict mobility scooter rules and speed limits in walkways to ensure pedestrian safety and prevent accidents.

4. Strengthen Friday Traffic Control: Increase traffic officer presence on Fridays, especially evenings, to ensure adherence to traffic rules and enhance road safety.

In conclusion, the predictive models for identifying accident severity offer a valuable tool to deploy traffic officials effectively in accident-prone regions. This approach enhances operational efficiency and upholds road safety.

Reference

Feng, M., Zheng, J., Ren, J. & Liu, Y. (2020) Towards big data analytics and mining for UK traffic accident analysis, visualization & prediction, *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*.

Haynes, S., Estin, P. C., Lazarevski, S., Soosay, M. & Kor, A.-L. (2019) Data analytics: Factors of traffic accidents in the uk, *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE.

Li, L., Shrestha, S. & Hu, G. (2017) Analysis of road traffic fatal accidents using data mining techniques, *2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE.

Gov.uk (2023) Checking what age you can drive . Available online: https://www.gov.uk/vehicles-can-drive?step-by-step-nav=e01e924b-9c7c-4c71-8241-66a575c2f61f. [Accessed 10/8/2023]