Accident Data Analysis Report: Navigating Towards Safety

1. Introduction

This report embarks on a comprehensive analysis of road accident data from 2020. The problem this analysis intends to solve is to reduce the number of road accidents and the severity of injuries caused by accidents, in case of occurrence, as accidents cannot be eliminated completely. Based on the findings, actionable insights would ensue for proactive, innovative, and AI-infused road safety measures.

2. Methodology

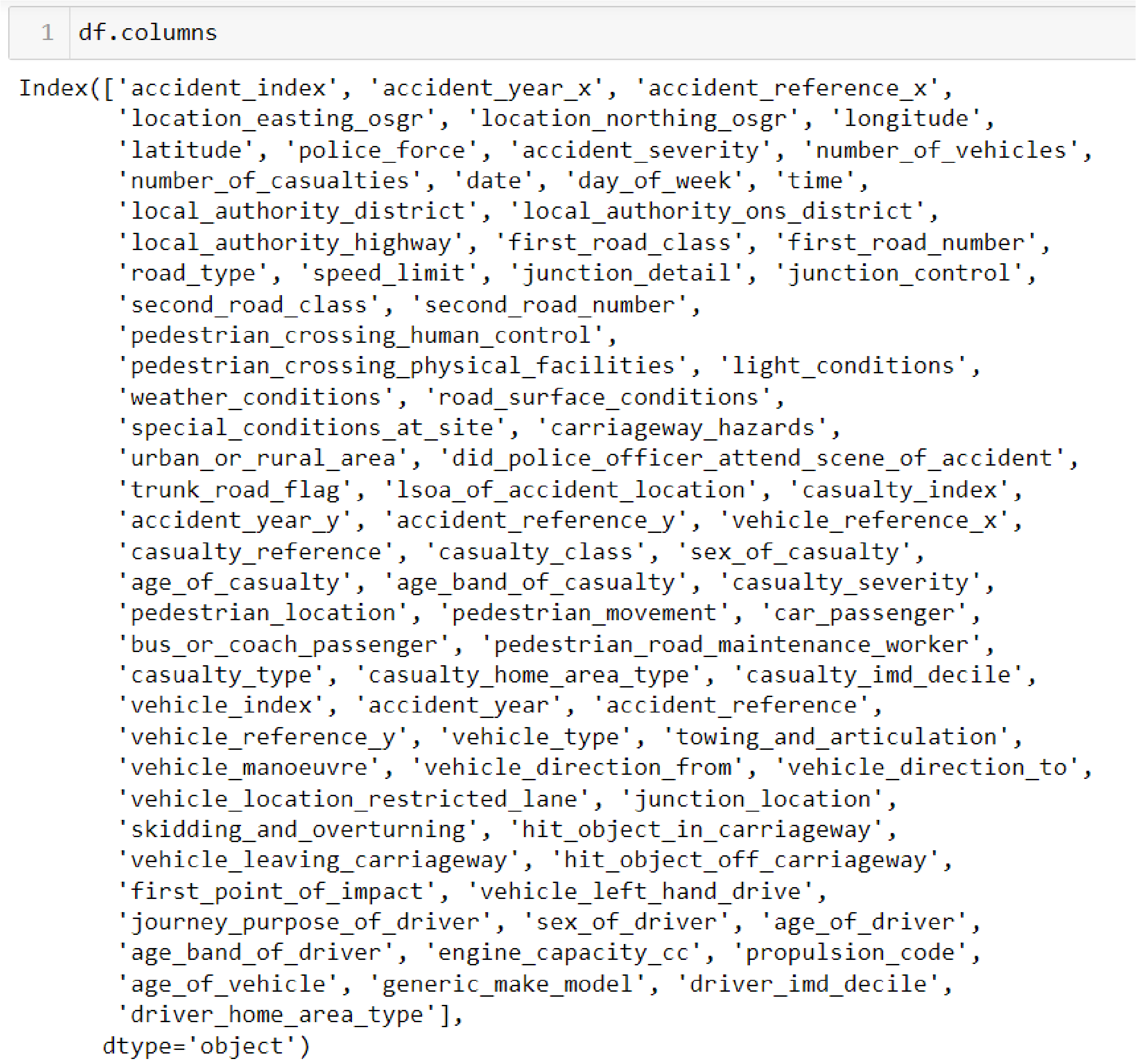
The analysis encompasses data preprocessing, outlier detection, feature selection, association rule mining, clustering, and predictive modeling.

- Data Preprocessing:
    - The initial step involved meticulous data cleaning to ensure the integrity of the dataset.
    - Nan Values in Longitude and Latitude columns were imputed using K-Neighbors Nearest (KNN) imputation, since it deals with distance preventing potential biases in subsequent analyses.
    - Numeric columns with -1 values were imputed using a Forward-Fill imputation, thereby maintaining the underlying data trend while addressing gaps.
    - Categorical columns with -1 values underwent Mode imputation, ensuring inferred values align with the prevailing category.
    - Erroneous codes, such as 99, were imputed using Stats20 as a reliable resource.

3. Exploratory Data Analysis:

Dataset Overview:

- The dataset encompasses records of 220,400 road traffic accidents.
- Analysis is done solely on 2020 data.
- The dataset encompasses 82 columns comprising variables associated with the accidents.
- It's important to note that the dataset has an imbalance.
- The dataset's quality is affected by its noisy nature, in form of missing values and outliers.



```
df.columns
```

```
Index(['accident_index', 'accident_year_x', 'accident_reference_x',
       'location_easting_osgr', 'location_northing_osgr', 'longitude',
       'latitude', 'police_force', 'accident_severity', 'number_of_vehicles',
       'number_of_casualties', 'date', 'day_of_week', 'time',
       'local_authority_district', 'local_authority_ons_district',
       'local_authority_highway', 'first_road_class', 'first_road_number',
       'road_type', 'speed_limit', 'junction_detail', 'junction_control',
       'second_road_class', 'second_road_number',
       'pedestrian_crossing_human_control',
       'pedestrian_crossing_physical_facilities', 'light_conditions',
       'weather_conditions', 'road_surface_conditions',
       'special_conditions_at_site', 'carriageway_hazards',
       'urban_or_rural_area', 'did_police_officer_attend_scene_of_accident',
       'trunk_road_flag', 'lsoa_of_accident_location', 'casualty_index',
       'accident_year_y', 'accident_reference_y', 'vehicle_reference_x',
       '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', 'casualty_imd_decile',
       'vehicle_index', 'accident_year', 'accident_reference',
       'vehicle_reference_y', 'vehicle_type', 'towing_and_articulation',
       'vehicle_manoeuvre', 'vehicle_direction_from', 'vehicle_direction_to',
       'vehicle_location_restricted_lane', 'junction_location',
       'skidding_and_overturning', 'hit_object_in_carriageway',
       'vehicle_leaving_carriageway', 'hit_object_off_carriageway',
       'first_point_of_impact', 'vehicle_left_hand_drive',
       'journey_purpose_of_driver', 'sex_of_driver', 'age_of_driver',
       'age_band_of_driver', 'engine_capacity_cc', 'propulsion_code',
       'age_of_vehicle', 'generic_make_model', 'driver_imd_decile',
       'driver_home_area_type'],
      dtype='object')
```

Figure 1: Summary of Accidents Dataset

3.1. Analysis of Accidents Based on Time and Day Patterns:

- Observations show an increase in accidents between Wednesdays and Friday at 8am - 5pm, due to work and school related travel.
- This trend is due to the bustling commute hours during weekdays and heightened social activities and parties on Fridays after work-hours.
- Sundays shows a lower accident count, indicative of a day of rest where people are indoors, hence reduced weekend traffic.

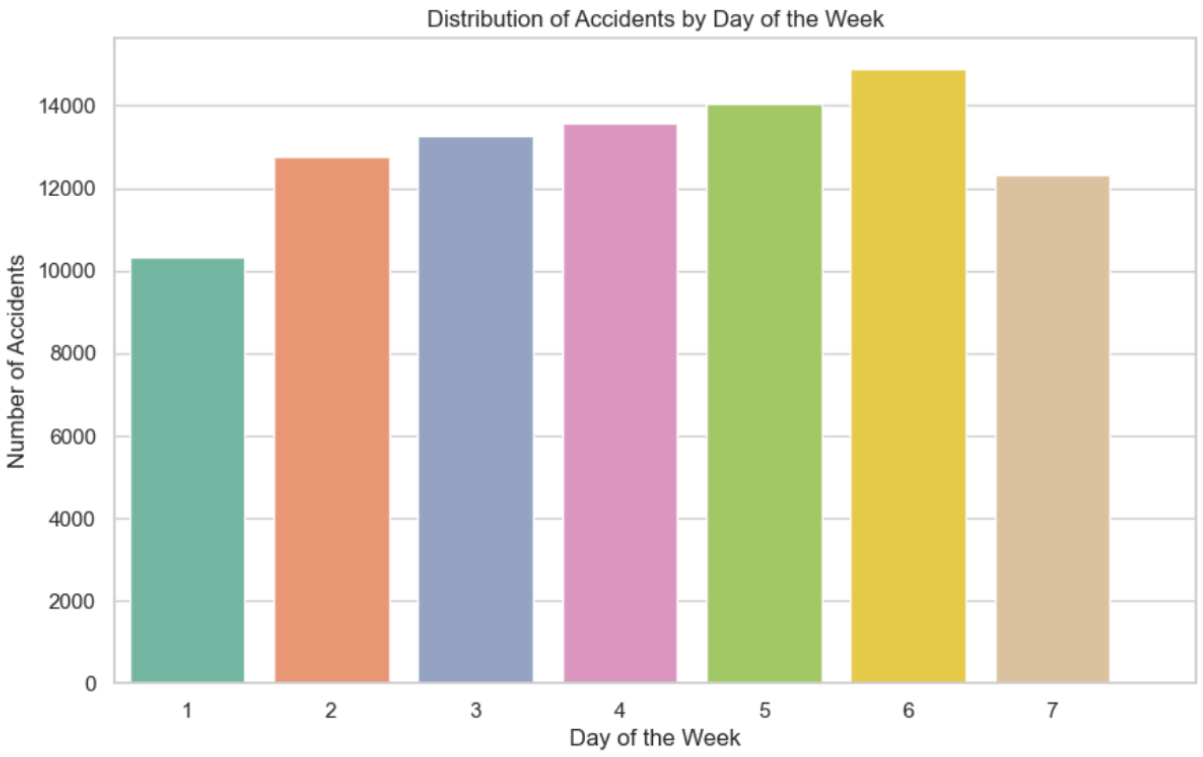


Figure 2: Number of accidents per day in a week

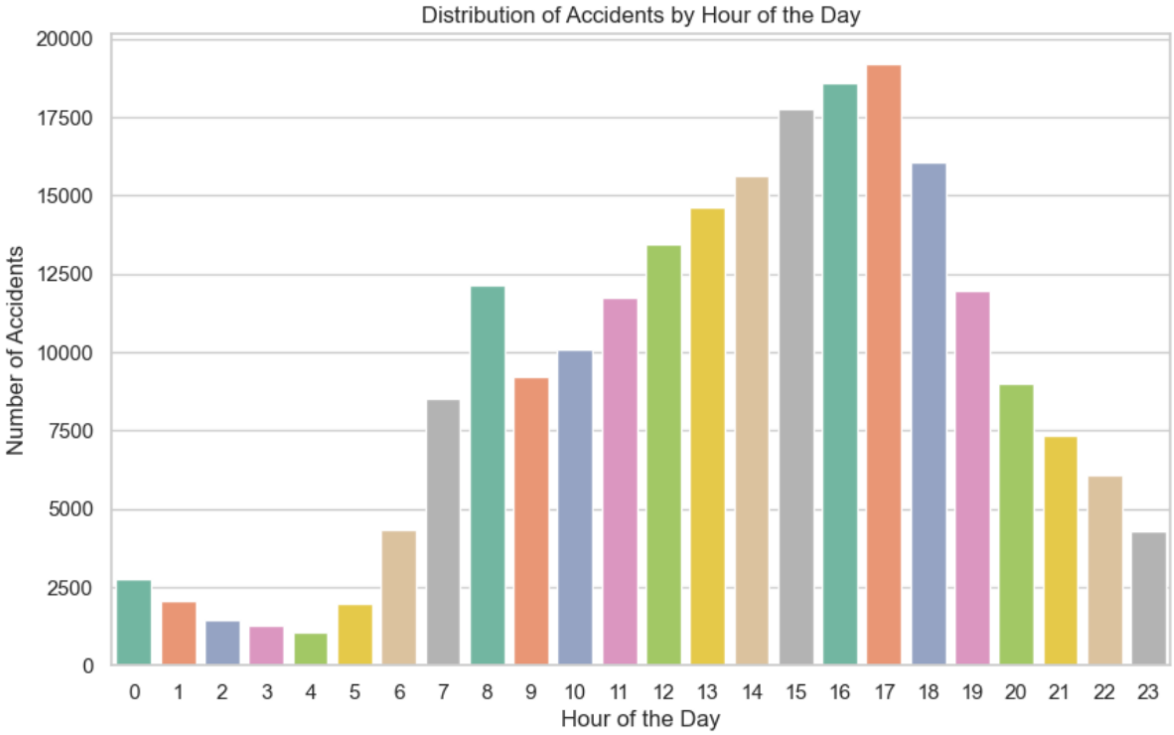


Figure 3: Number of accidents per hour in a day

3.2. Motorbike-Related Accidents Analysis:

Engine Capacity and Accidents:

As shown in figures 4,6,8. Motorbikes with engine capacities up to 500cc, as well as those above 500cc, emerge with more numbers of accidents than 125cc and under. This analysis supports the findings of a previous study that the risk of an injury crash increases with increasing engine capacity (Langley et al., 2000)

Temporal Patterns:

- Peak accident hours, such as 8am and 5pm also occurs with motorbikes indicative of rush hours.
- Motorbike accidents exhibit a pronounced surge on Fridays and Saturdays for under 500cc as they are most time drunk-driving as it is the weekend and with high speed. Over 500cc occur on Sundays with more accidents.

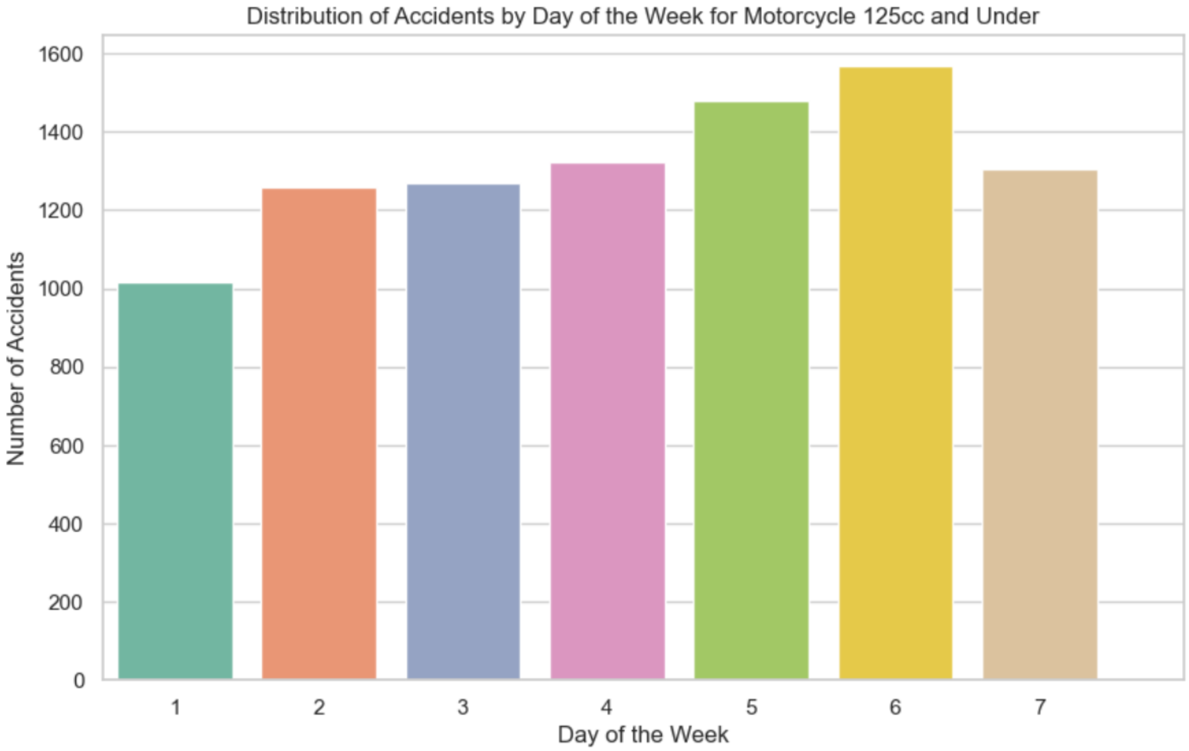


Figure 4:125cc and under accidents per day in a week

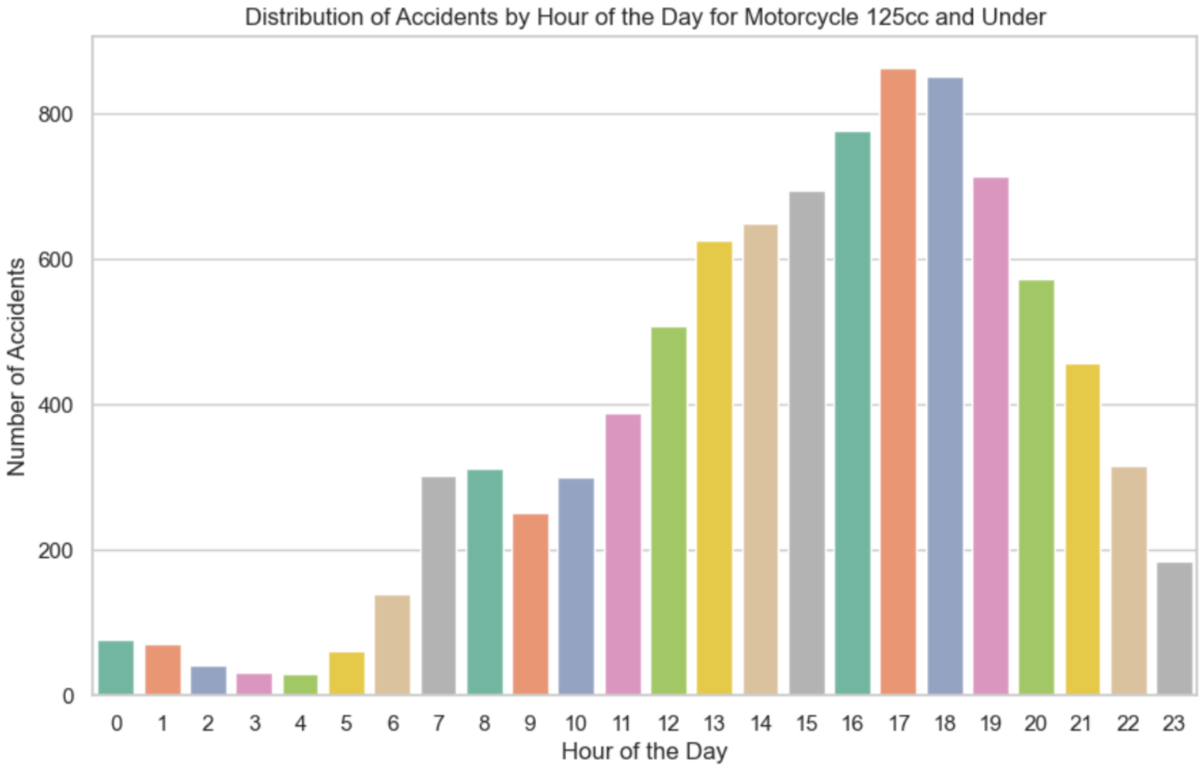


Figure 5:125cc and under accidents per hour in a day

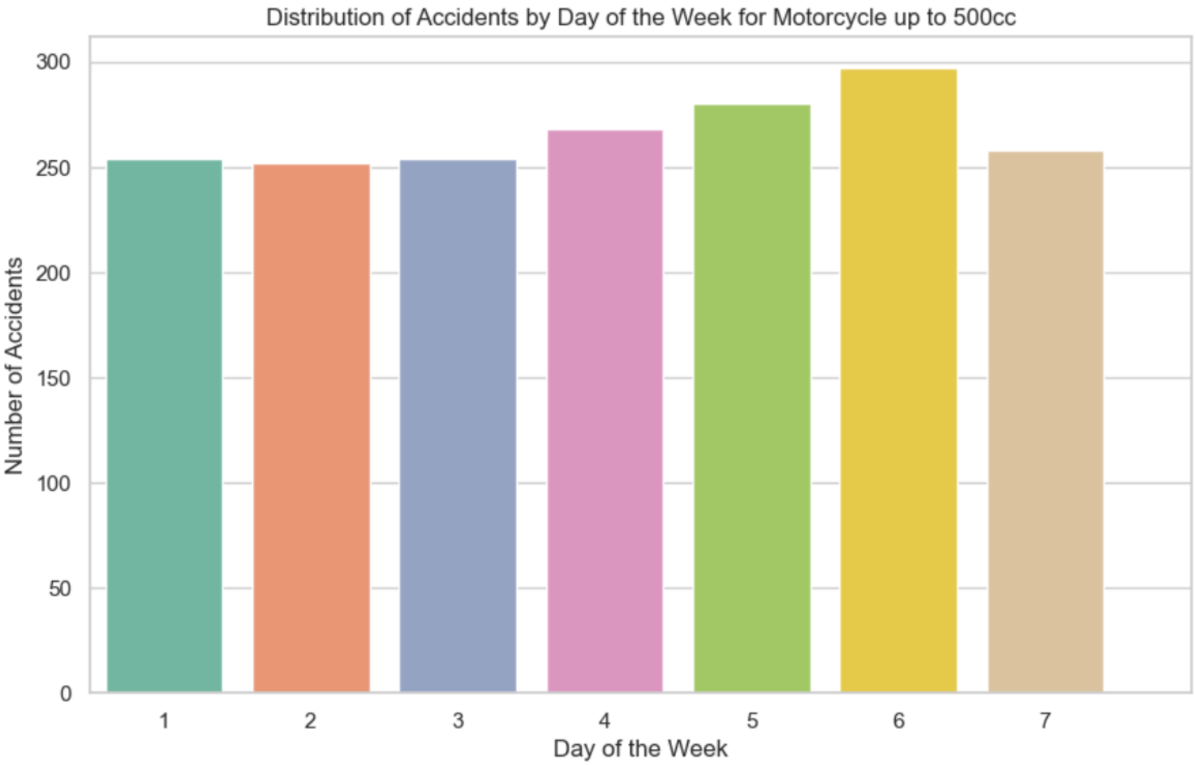


Figure 6: Up to 500cc accidents per day in a week

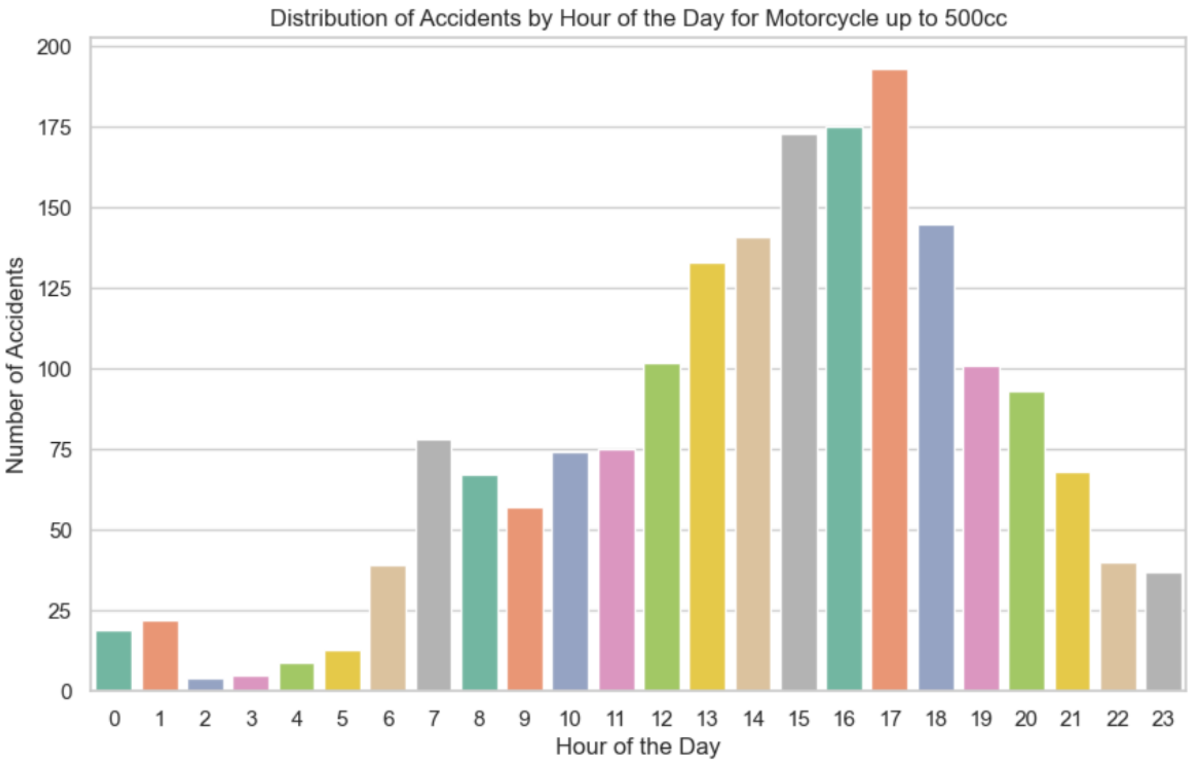


Figure 7: Up to 500cc accidents per hour in a day

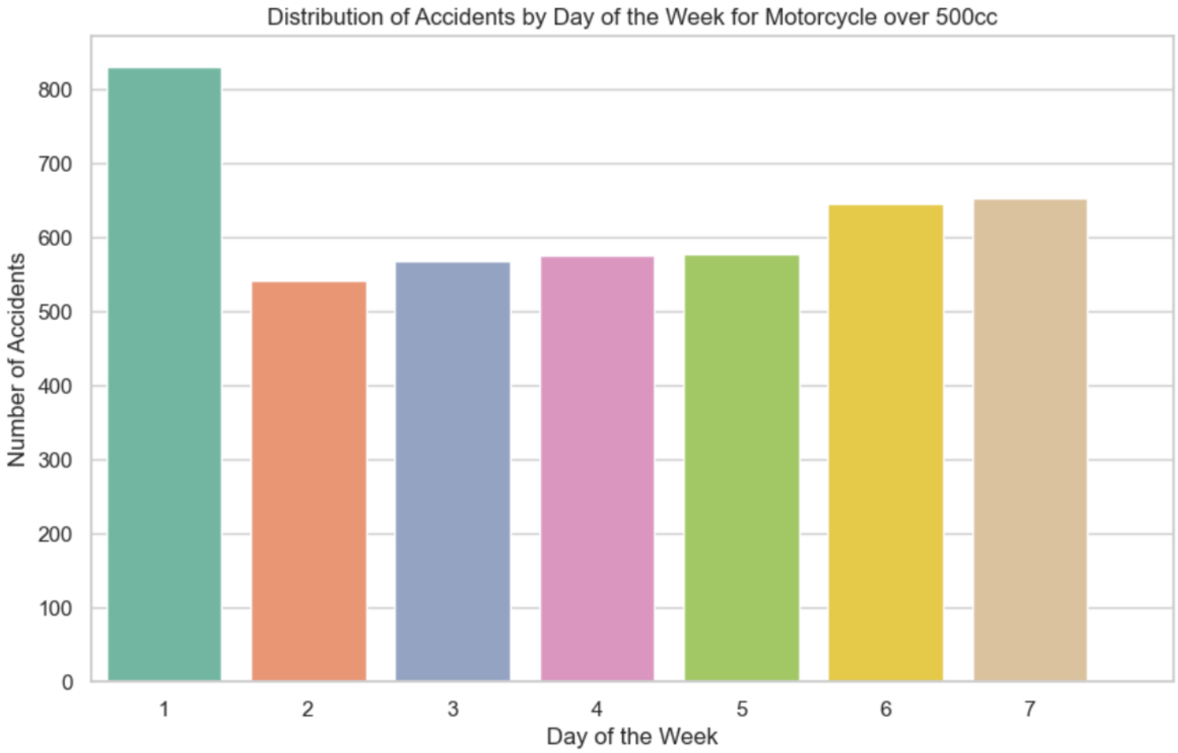


Figure 8: Over 500cc accidents per day in a week

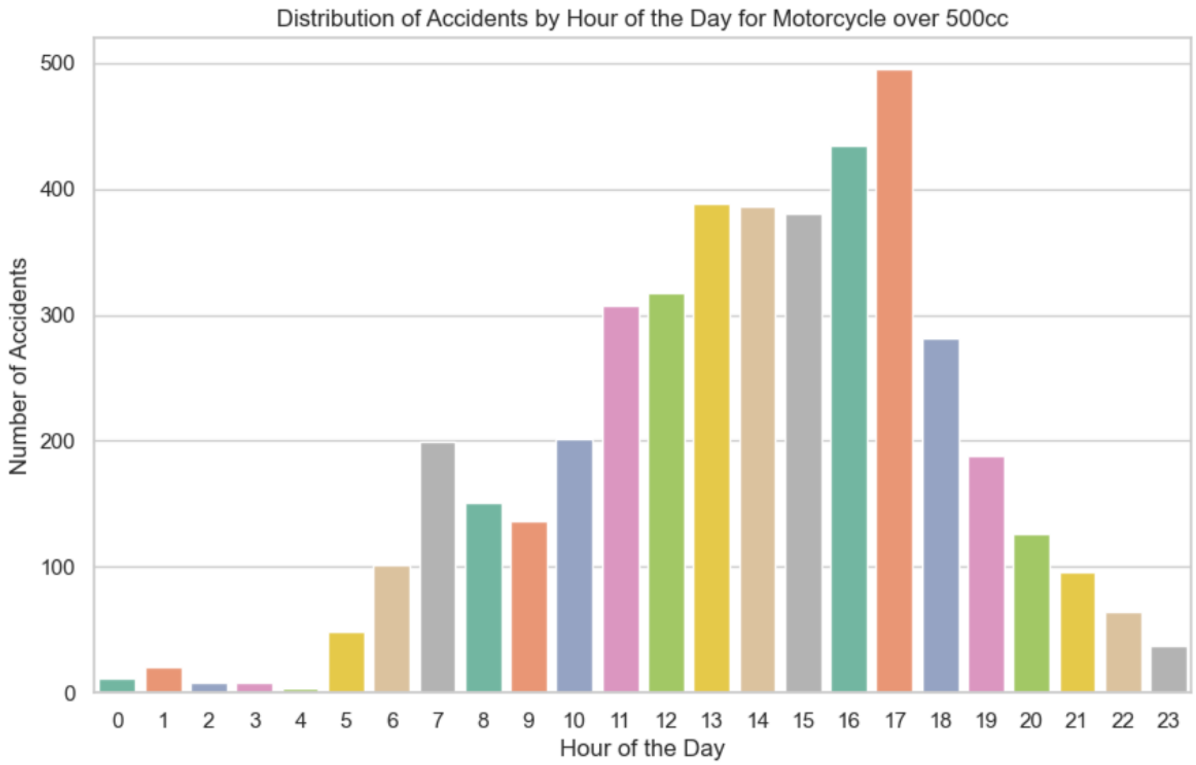


Figure 9: Over 500cc accidents per hour in a day

3.3. Pedestrian Accident by Day of The Week, Hour:

Figure 10 & 11 shows that most pedestrian accidents occurred during the weekday, with a higher number on Saturday around 8am and 3pm. The weekday high number is attributed to people walking to and from bus stops either to go to school or work would make up most of the casualties during this time. The high number on Saturday is attributed to people going for social activities. Furthermore, (Jang et al.,2013) have demonstrated that the severity of pedestrian crashes escalates as the weekend nears.

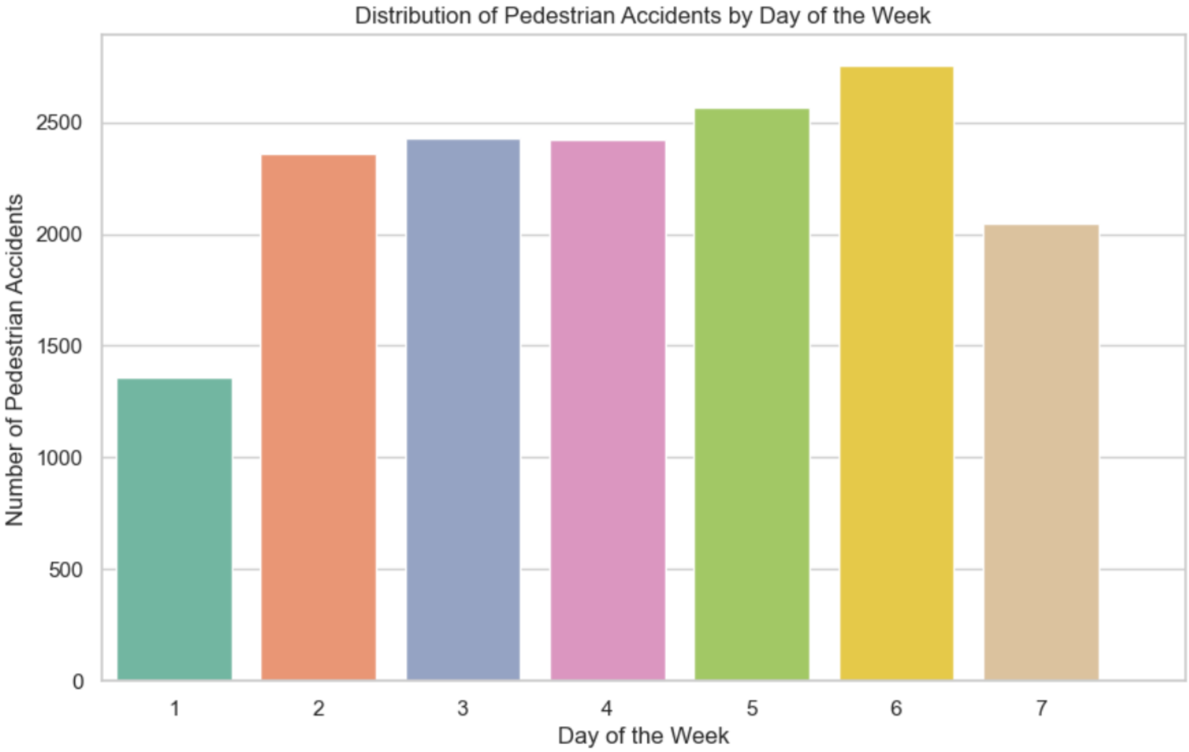


Figure 10: Pedestrian accidents per day in a week

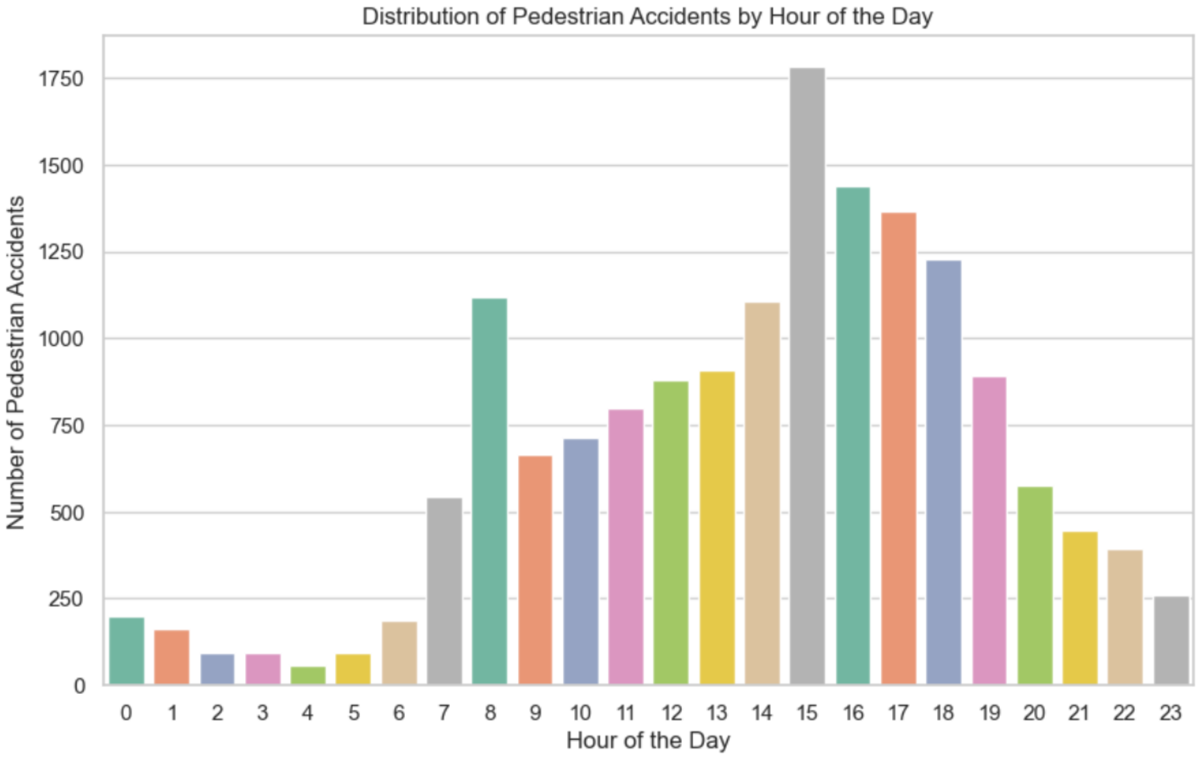


Figure 11: Pedestrian accidents per hour in a day

4. Association Pattern Mining:

The most significant features extracted from K-Best are the selected variables used here. A lift value >1 and confidence >0.5 was used.

<u>Strongest Association Rules:</u>

Rule 1:

Association: `Urban Area => Speed Limit: 30, Accident Severity: Slight`

- Support: 37.84%, Confidence: 86.19%, Lift: 1.3418, Zhang's Metric: 0.4541

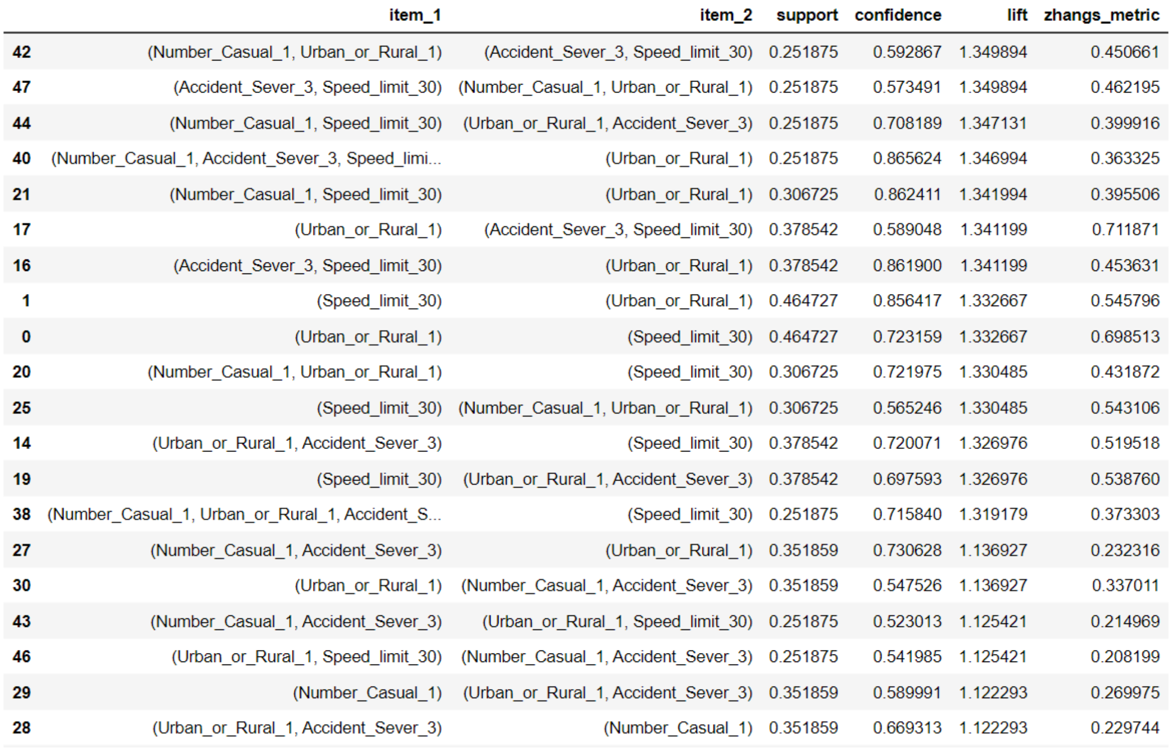
This rule suggests that in urban areas, there is a strong association between accidents with a speed limit of 30 and a slight level of accident severity. First, urban areas are typically more densely populated than other areas, hence many urban areas have a speed limit of 30. An accident severity that occurs at the moderate speed limit of 30 would be slight as against an occurrence at a higher speed limit.

Rule 2:

Association: `Urban Area, 1 Casualty => Speed Limit: 30, Accident Severity: Slight`

- Support: 25.15%, Confidence: 59.28%, Lift: 1.3503, Zhang's Metric: 0.4506

This rule highlights that in both urban areas, accidents with a speed limit of 30 and involving one casualty are more likely to result in a slight level of accident severity. This association would most likely have fewer casualties.



	item_1	item_2	support	confidence	lift	zhangs_metric
42	(Number_Casual_1, Urban_or_Rural_1)	(Accident_Sever_3, Speed_limit_30)	0.251875	0.592867	1.349894	0.450661
47	(Accident_Sever_3, Speed_limit_30)	(Number_Casual_1, Urban_or_Rural_1)	0.251875	0.573491	1.349894	0.462195
44	(Number_Casual_1, Speed_limit_30)	(Urban_or_Rural_1, Accident_Sever_3)	0.251875	0.708189	1.347131	0.399916
40	(Number_Casual_1, Accident_Sever_3, Speed_limi...	(Urban_or_Rural_1)	0.251875	0.865624	1.346994	0.363325
21	(Number_Casual_1, Speed_limit_30)	(Urban_or_Rural_1)	0.306725	0.862411	1.341994	0.395506
17	(Urban_or_Rural_1)	(Accident_Sever_3, Speed_limit_30)	0.378542	0.589048	1.341199	0.711871
16	(Accident_Sever_3, Speed_limit_30)	(Urban_or_Rural_1)	0.378542	0.861900	1.341199	0.453631
1	(Speed_limit_30)	(Urban_or_Rural_1)	0.464727	0.856417	1.332667	0.545796
0	(Urban_or_Rural_1)	(Speed_limit_30)	0.464727	0.723159	1.332667	0.698513
20	(Number_Casual_1, Urban_or_Rural_1)	(Speed_limit_30)	0.306725	0.721975	1.330485	0.431872
25	(Speed_limit_30)	(Number_Casual_1, Urban_or_Rural_1)	0.306725	0.565246	1.330485	0.543106
14	(Urban_or_Rural_1, Accident_Sever_3)	(Speed_limit_30)	0.378542	0.720071	1.326976	0.519518
19	(Speed_limit_30)	(Urban_or_Rural_1, Accident_Sever_3)	0.378542	0.697593	1.326976	0.538760
38	(Number_Casual_1, Urban_or_Rural_1, Accident_S...	(Speed_limit_30)	0.251875	0.715840	1.319179	0.373303
27	(Number_Casual_1, Accident_Sever_3)	(Urban_or_Rural_1)	0.351859	0.730628	1.136927	0.232316
30	(Urban_or_Rural_1)	(Number_Casual_1, Accident_Sever_3)	0.351859	0.547526	1.136927	0.337011
43	(Number_Casual_1, Accident_Sever_3)	(Urban_or_Rural_1, Speed_limit_30)	0.251875	0.523013	1.125421	0.214969
46	(Urban_or_Rural_1, Speed_limit_30)	(Number_Casual_1, Accident_Sever_3)	0.251875	0.541985	1.125421	0.208199
29	(Number_Casual_1)	(Urban_or_Rural_1, Accident_Sever_3)	0.351859	0.589991	1.122293	0.269975
28	(Urban_or_Rural_1, Accident_Sever_3)	(Number_Casual_1)	0.351859	0.669313	1.122293	0.229744

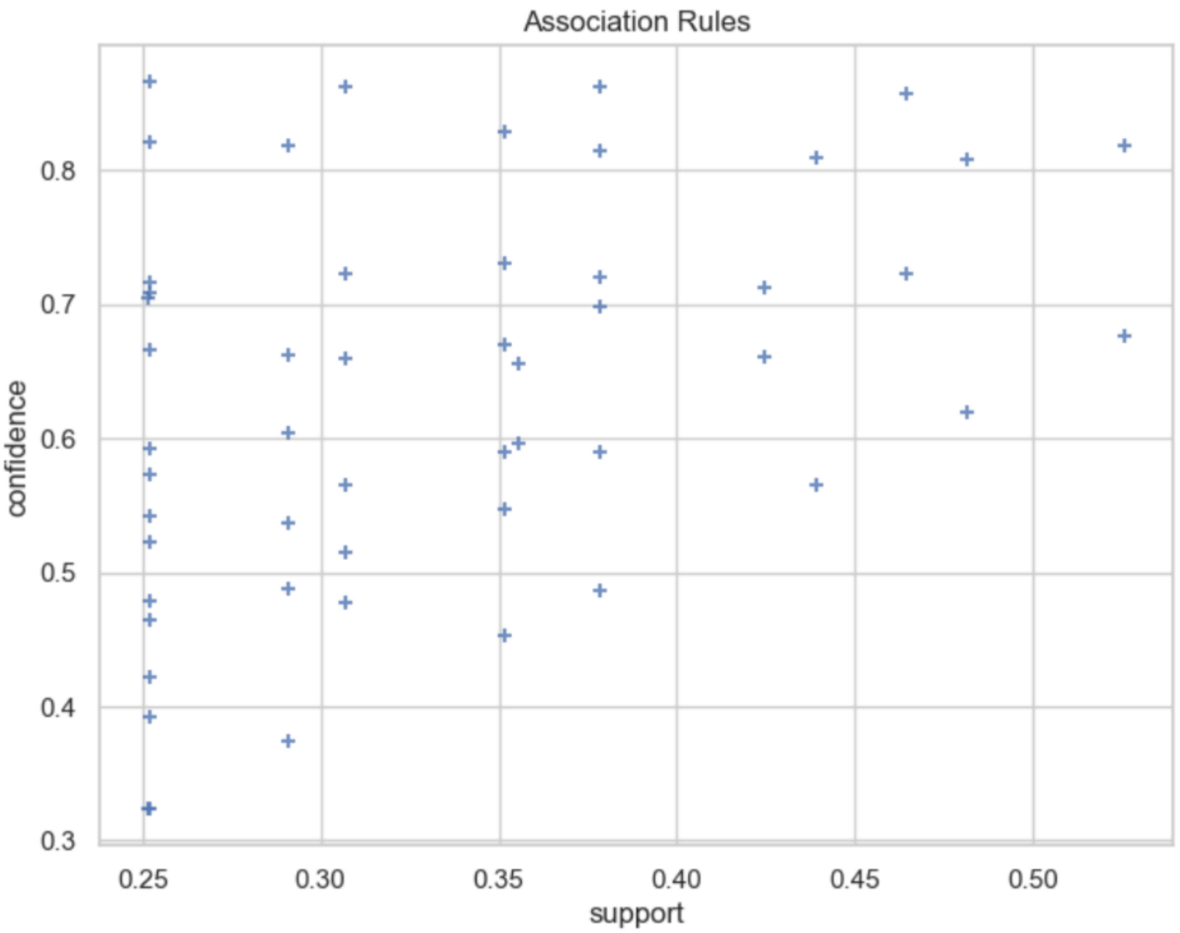


Figure 12: Association Pattern Mining

5. Optimal Cluster Identification:

Using the elbow method, an optimal number of (5) clusters was utilized to unearth a compact, well-separated, and balanced clusters in the regions.

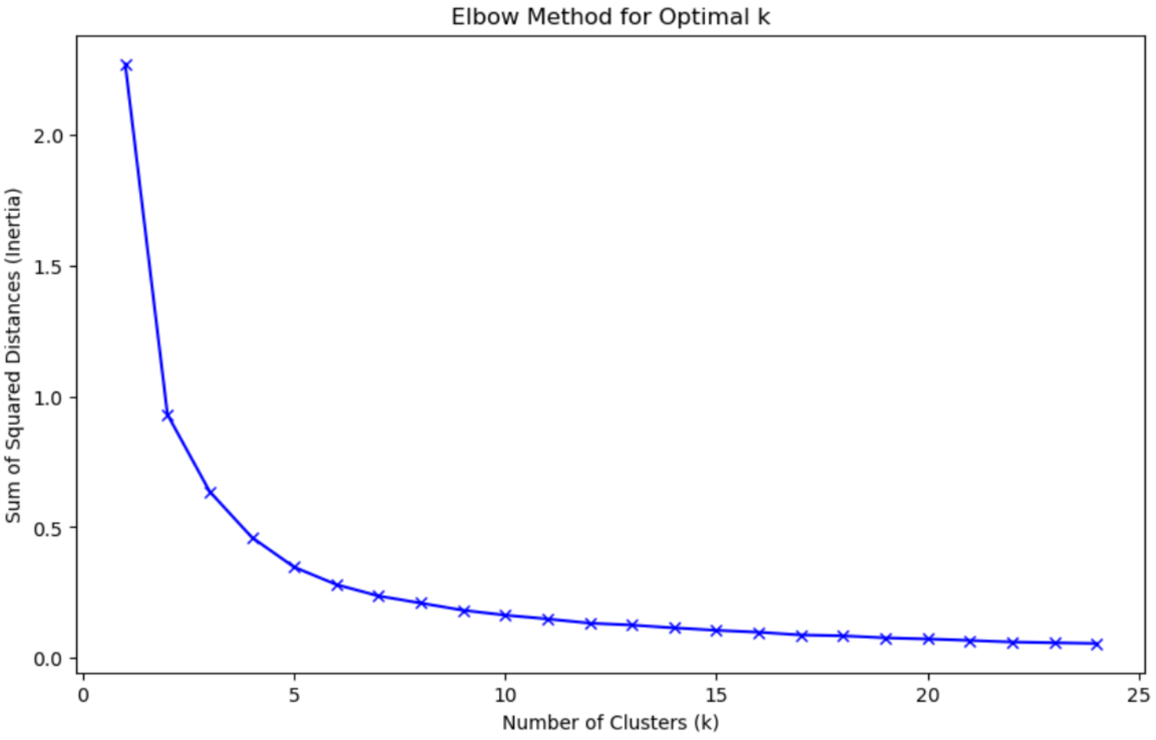


Figure 13: Elbow method for optimal clusters

5.1. Accident Cluster in Kingston Upon Hull:

As shown in figure 14, the largest cluster is in the central part of Kingston upon Hull. It is a busy and high traffic area with many businesses and the train station. The centroid of the cluster in the center of the city is located near the intersection of Anlaby Road and Hessle and this intersection can be confusing for drivers, especially those who are unfamiliar with the area.

Another cluster is near Marfleet which is home to several industrial businesses, which implies more heavy-duty trucks leading to traffic congestion.

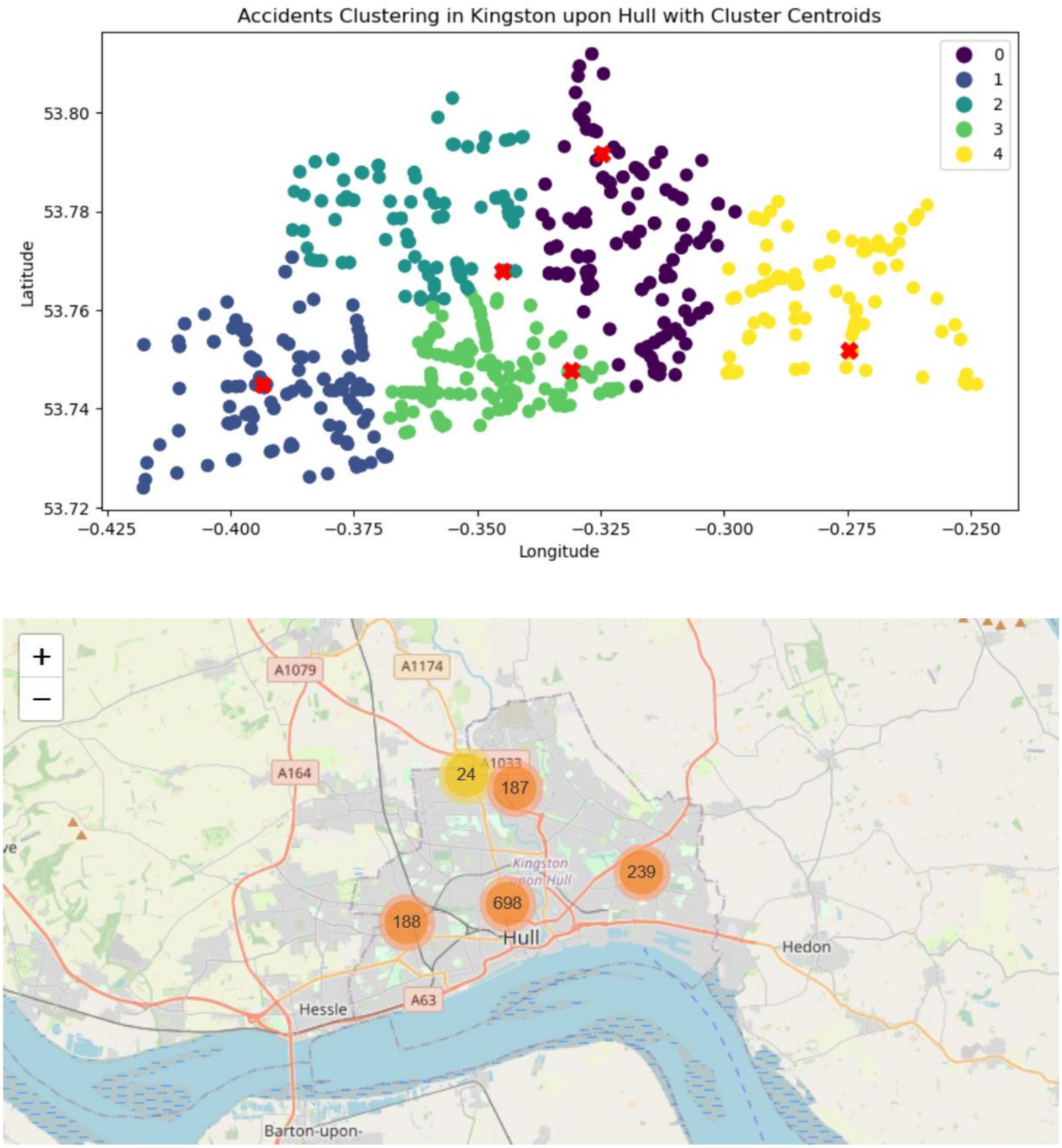
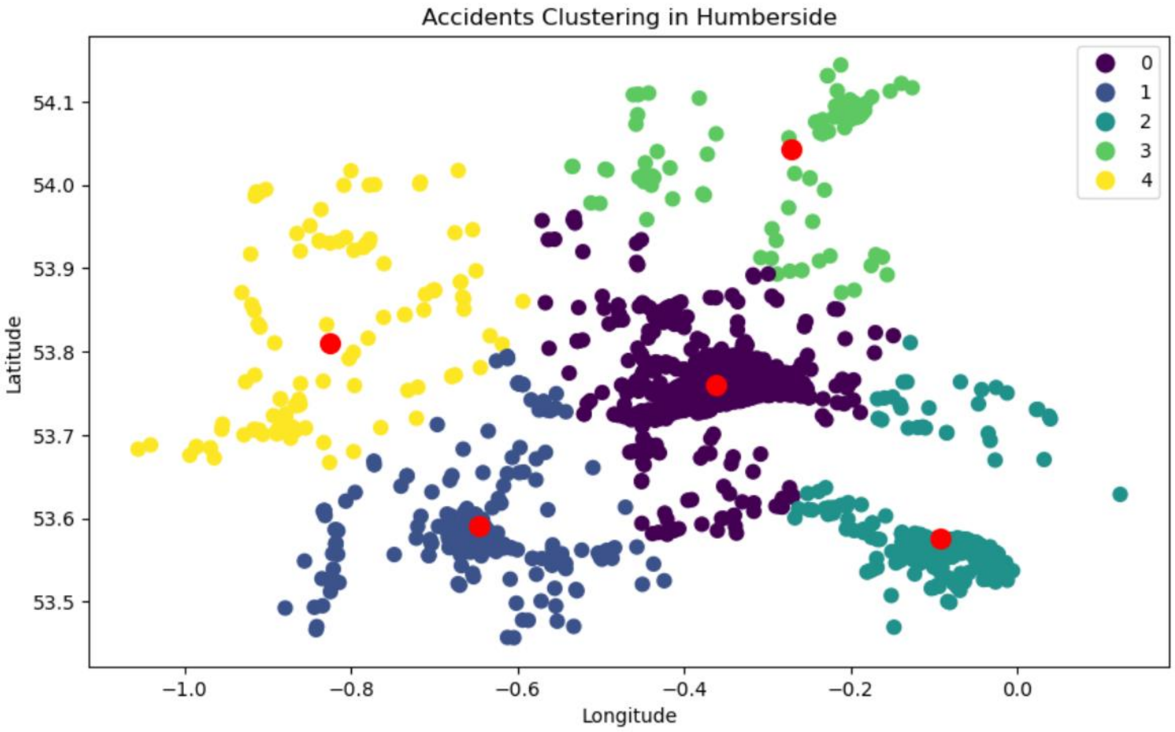


Figure 14: Kingston upon Hull Accident Cluster and Map

5.2. Accident Cluster in Humberside Region:

As shown in figure, there is a large cluster of accidents near Hull close to the Humber River. The area is rural, with limited visibility due to fog and mist from the water body causing accidents due to poor weather condition. The Humber bridge attracts a lot of tourists, resulting to accidents involving tourists in this area.

Another large cluster of accidents occurred near Scunthorpe. This cluster is located on the A180, which is a major road that which can create traffic congestion and merging hazards responsible for accidents.



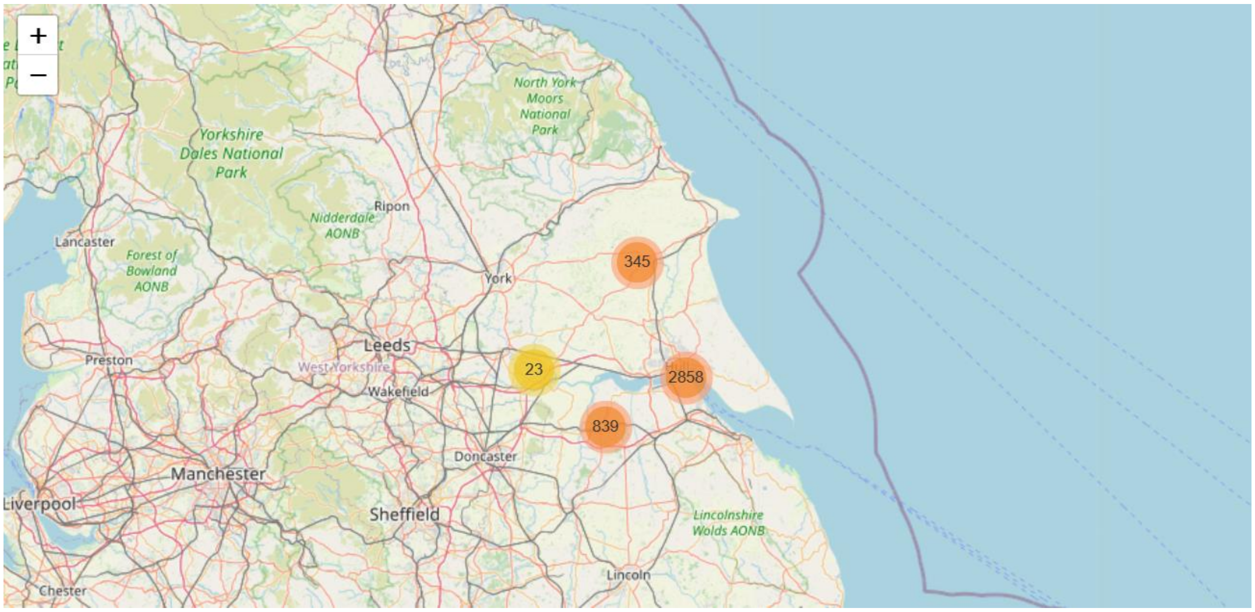


Figure 15: Humberside Accident Cluster and Map

5.3. Accident Cluster in East-Riding Region:

There is a large cluster of accidents near Beverley on the East Riding map. This cluster is located on the A165, which is rural, with limited visibility. The intersection of the A165 and the A614 is also a complex intersection with a high volume of traffic which can cause accidents.

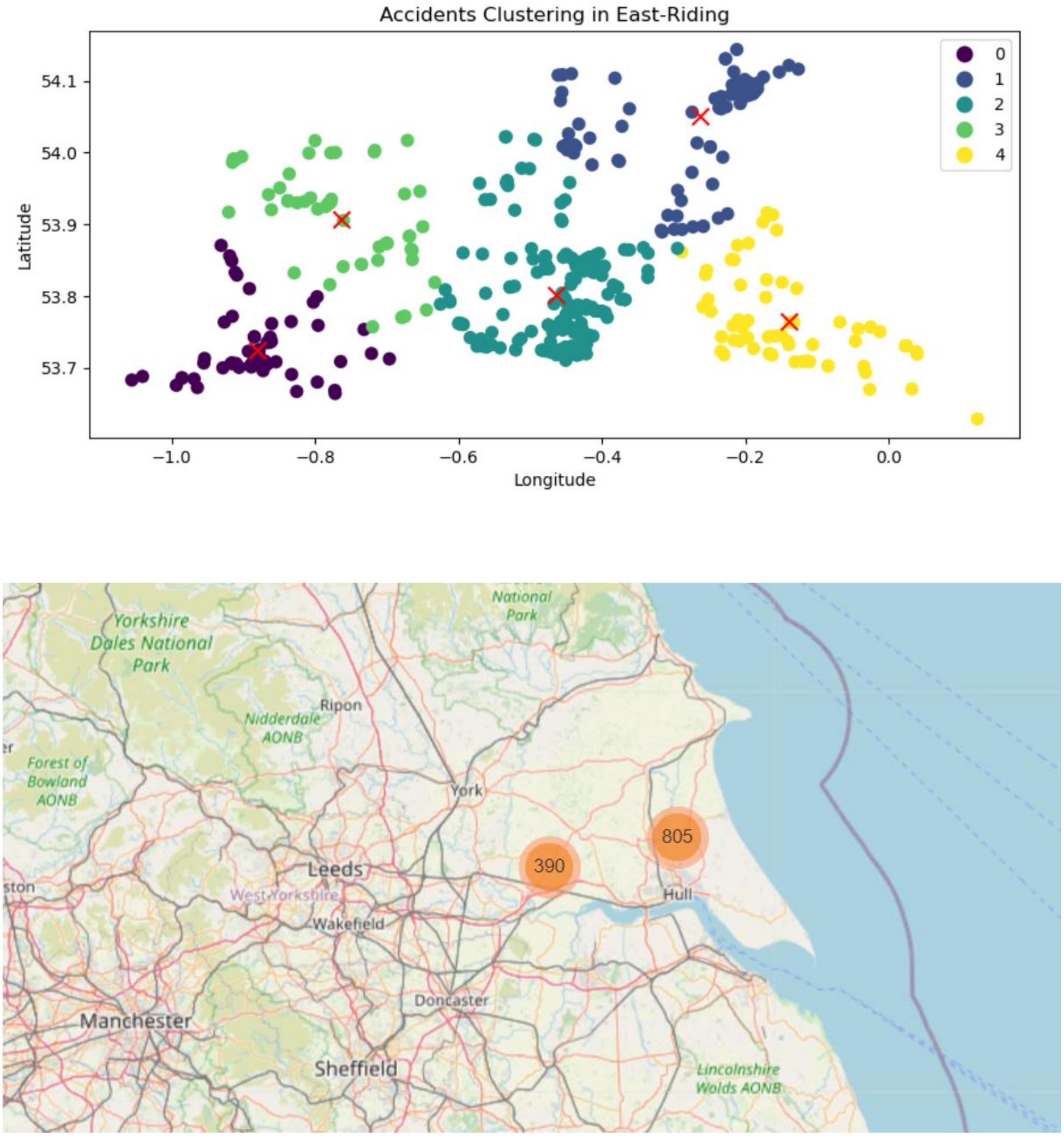


Figure 16: East Riding of Yorkshire Accident Cluster and Map

6. Outlier Detection & Handling:

This section provides an account of the outlier detection process, which includes initial visualization, Skew Analysis, Grubbs' test for isolated data points, and the application of the Isolation Forest algorithm for comprehensive analysis. Figure 17 is a distribution based on outliers identified by Isolated Forest Algorithm

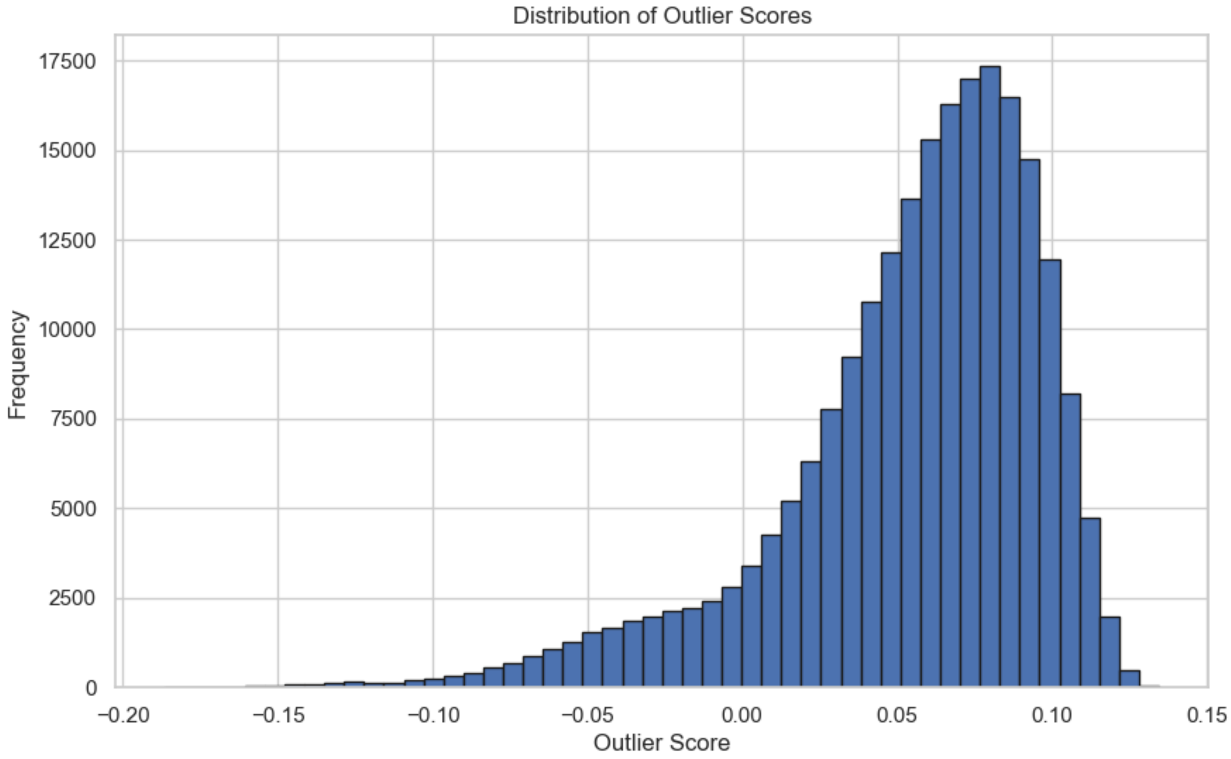


Figure 17: Outlier Distribution Before Removal

Age of Driver Outlier:

Data points showed where drivers below the age of 16years drove cars, trucks, and were involved in accidents. The age for driving is 17 years and for a person to drive at age 16 years, he must have obtained PIP (Gov.uk). Hence, this motivated my decision to leave the data of 16 years old drivers. Those below 16 years were dropped as it is not possible for Driver's license to be issued to an underage. Some were as low as age 10years old driving a Ford Fiesta car. There are some drivers up to 100 years old but would not be removed. Drivers above 70 years can renew their driver's license every three years (Gov.uk).

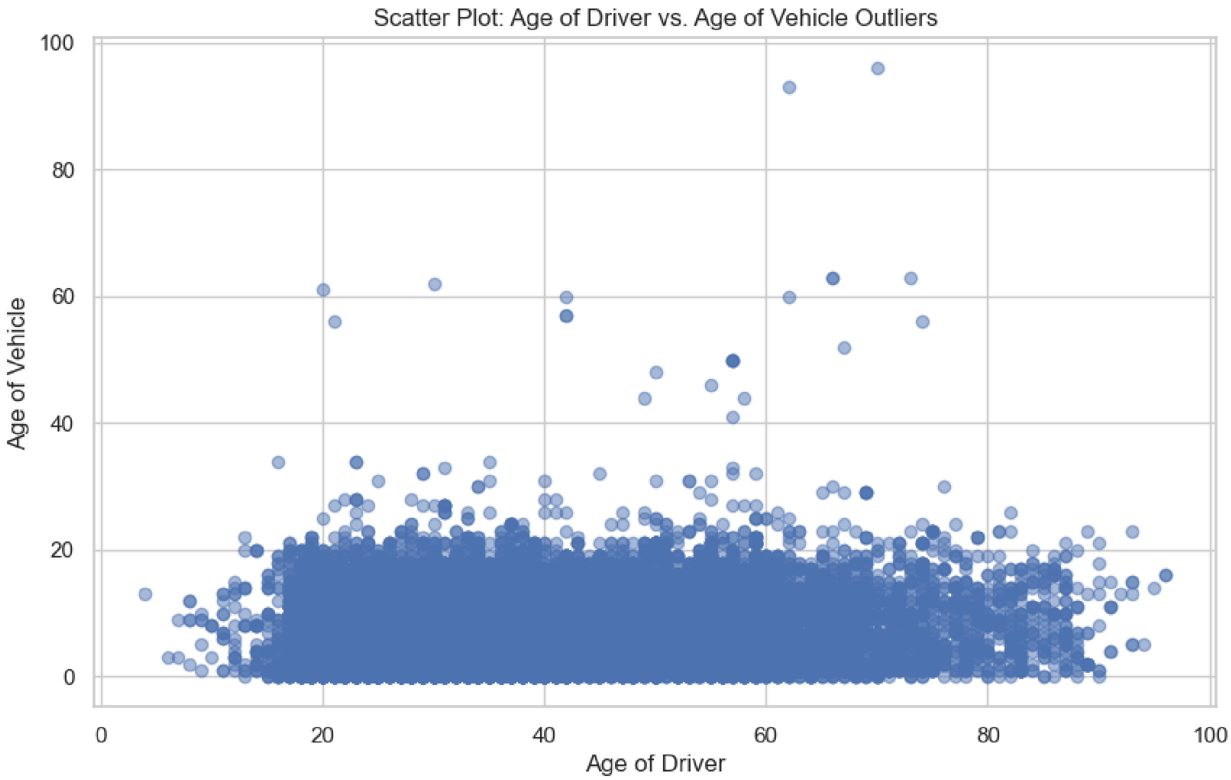


Figure 18: Age of Driver & vehicle Outlier

Age of Vehicle Outliers:

Scatter plots and box plots revealed that vehicles older than 80 years were recorded as being involved in accidents. This finding is unusual and was dropped. But the vehicles between 20years and lesser than 80 years were left as they did not pose as distinct outliers in the plot. Vehicles older than 40 years would get MOT exemption and be termed as classic vehicles and must ensure no substantial changes has been made to the vehicle parts in the last 30 years (Gov.uk).

Engine Capacity cc Outliers:

An entry for engine capacity of Ford Fiesta was above 17500 cc. There was also a Vauxhall vehicle of above 17500 cc. The average engine capacity cc of Vauxhall is 2275cc (Enginert.com). Also, the engine capacity cc of Ford Fiesta falls within 999 cc and 1596 cc (Cars-Data.com). The entry of 17500 cc is significantly higher than the expected engine capacity for these models. Instead of dropping this data, it would be imputed with the mode engine capacity cc of other ford fiesta cars and Vauxhall car respectively.

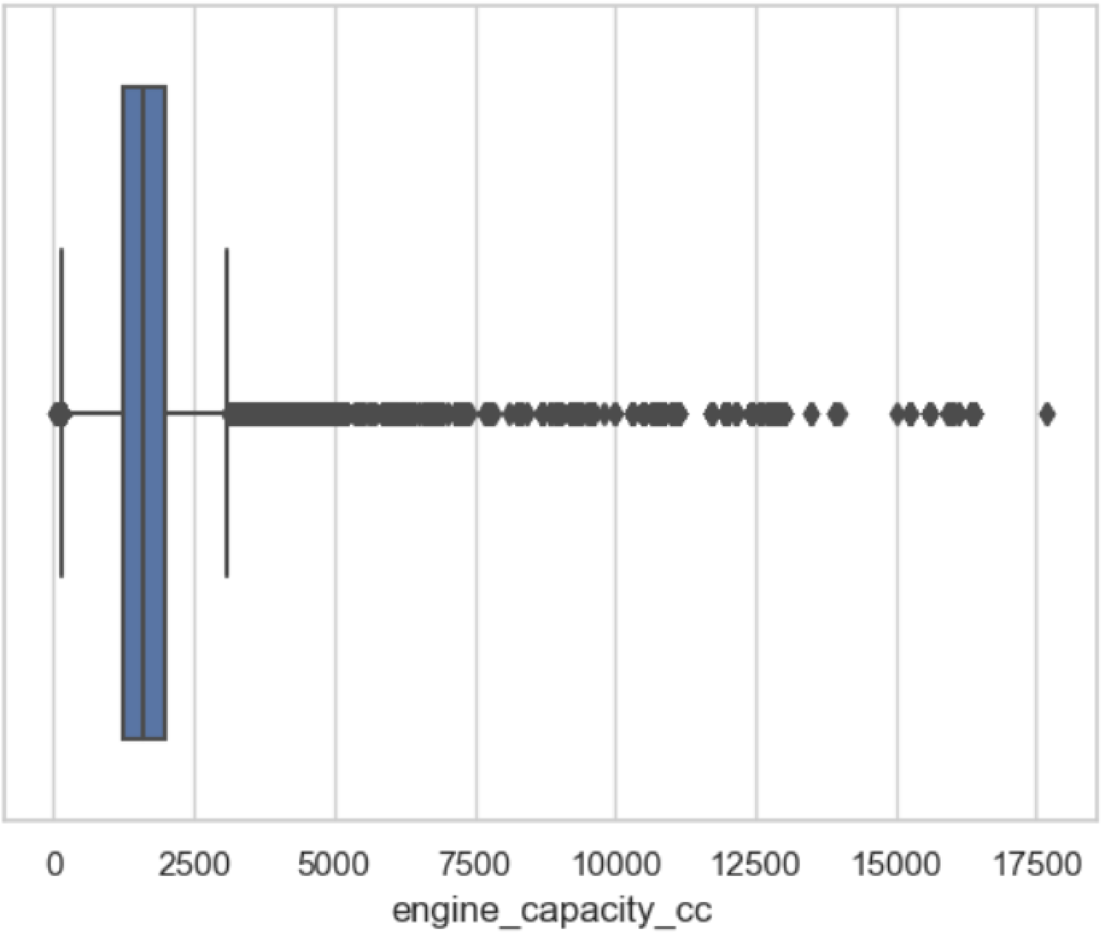


Figure 19: Engine Capacity cc Outlier

6.1. Choice of Outlier Detection Algorithm:

Isolation Forest algorithm was chosen as technique suitable for multivariate datasets with high dimensionality, to check the entire dataset for outliers. Grubb’s test was further used univariately to inspect each column.

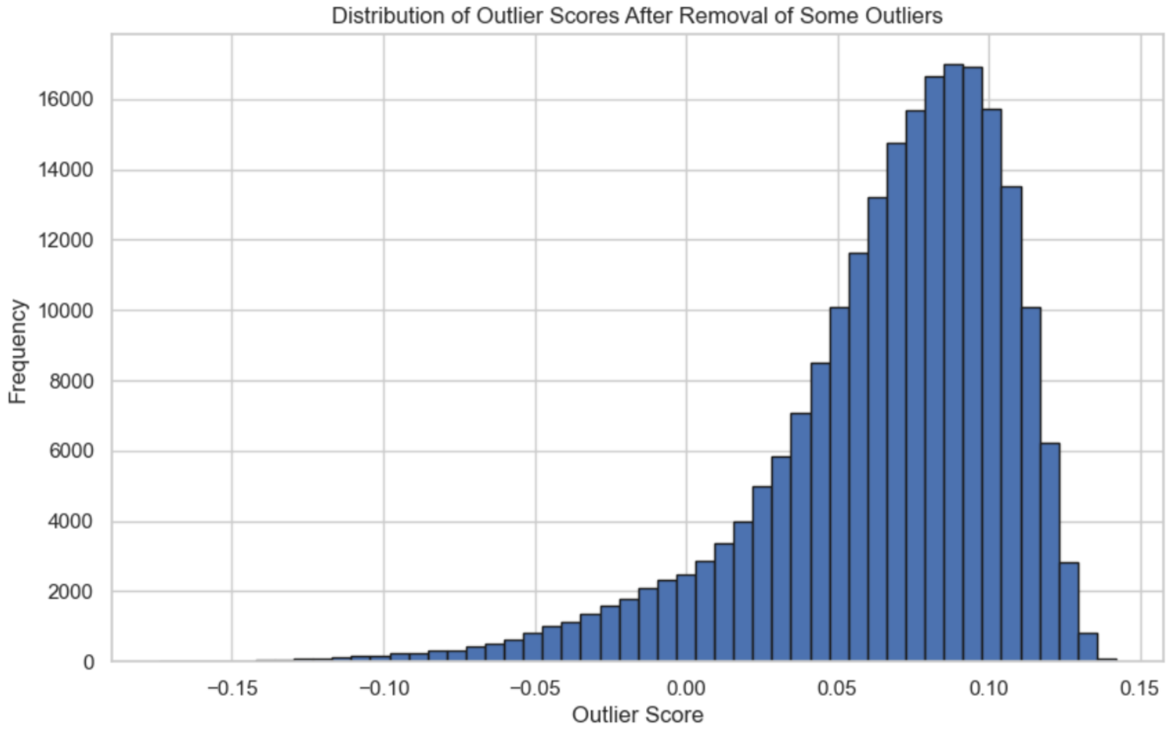


Figure 20: Outlier Distribution After Removal

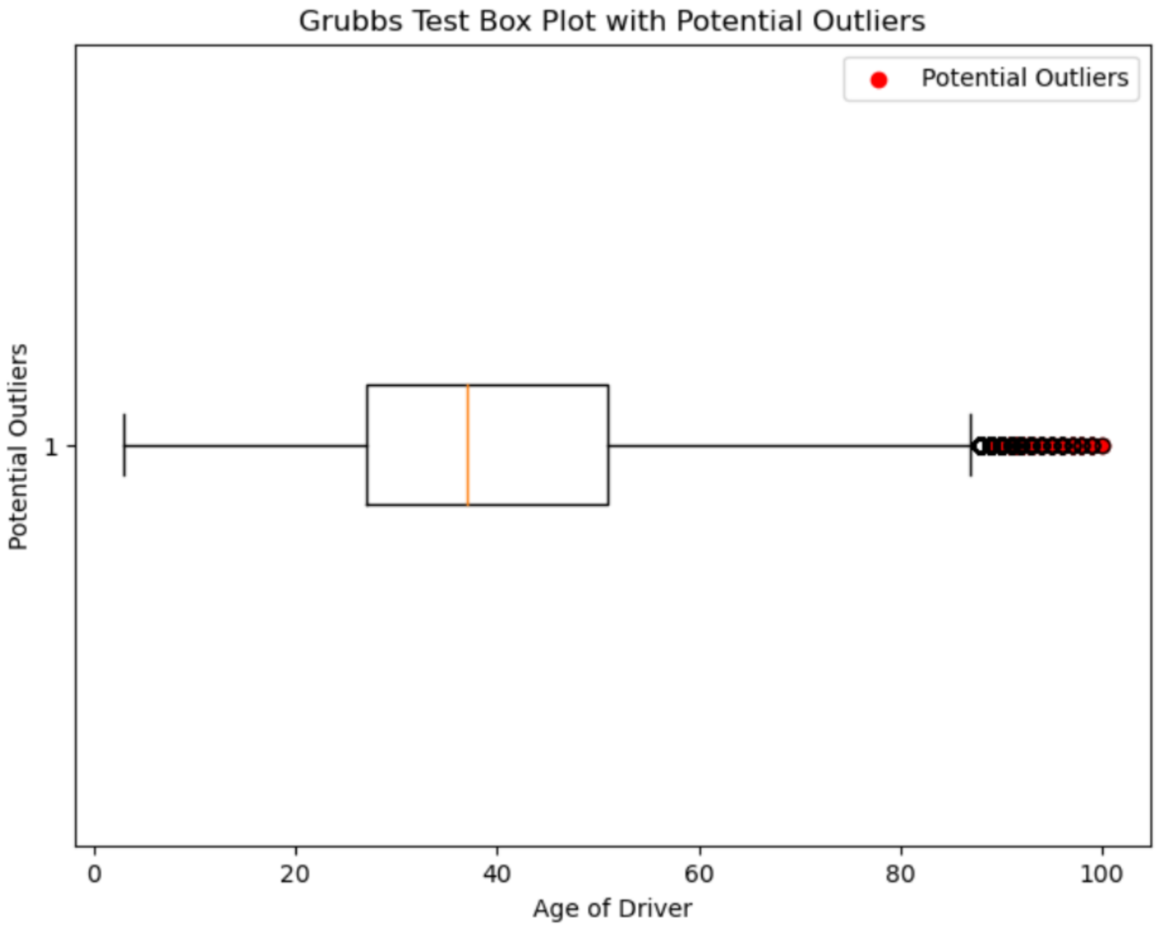


Figure 21: Grubb's Test Outlier Detection

7. Predictions:

K-best selected features were used to carry out the predictions. There was a significant change in the performance metrics of the classification before and after outliers were removed as seen in the figure.

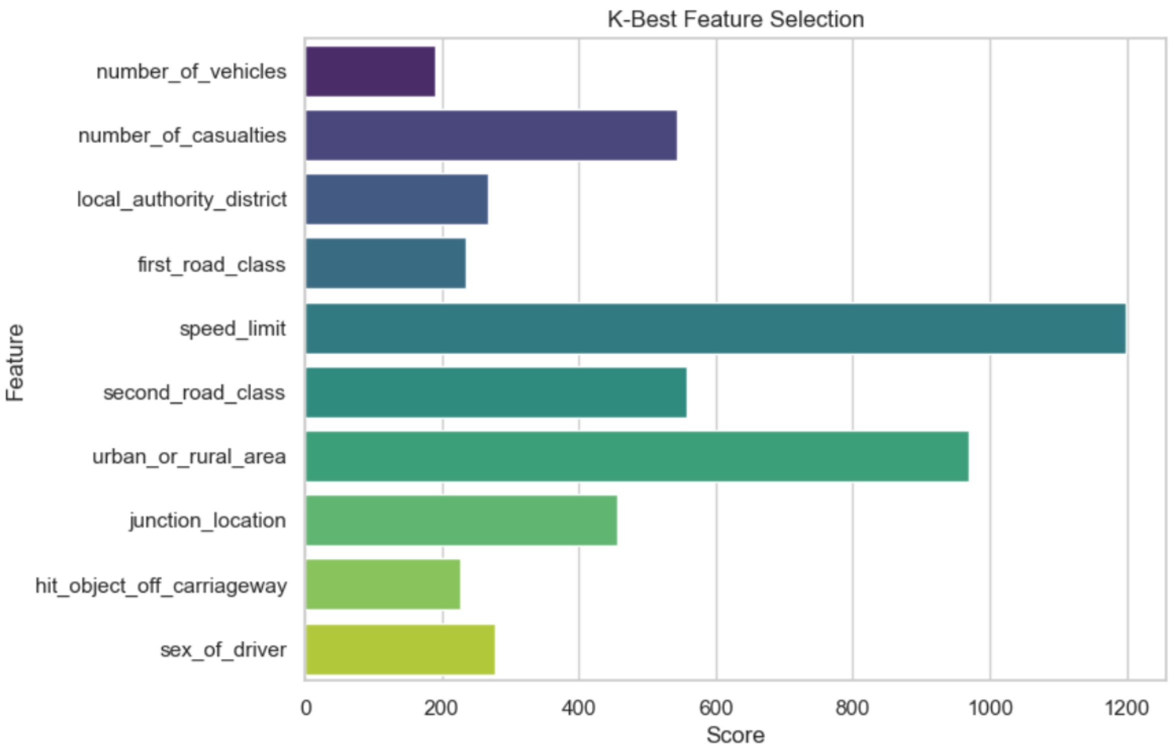
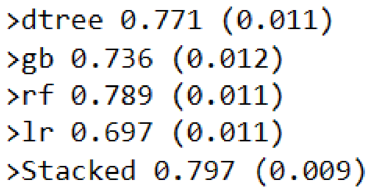
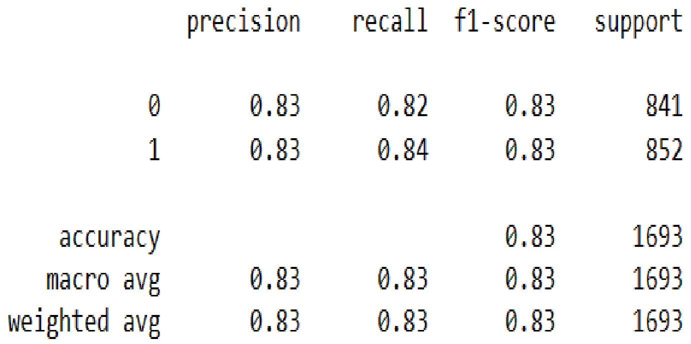


Figure 22: K-Best Feature Selection

Before Outlier Removal:

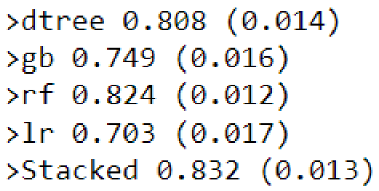


```
>dtree 0.771 (0.011)
>gb 0.736 (0.012)
>rf 0.789 (0.011)
>lr 0.697 (0.011)
>Stacked 0.797 (0.009)
```

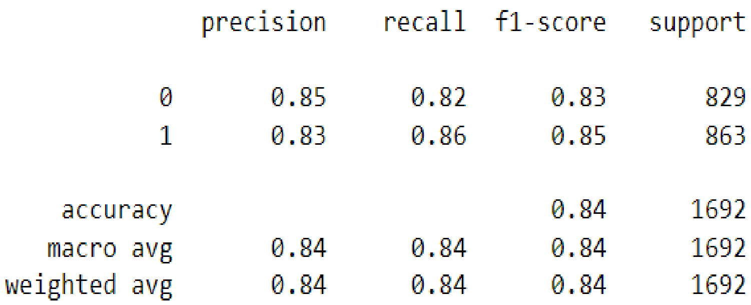


```
              precision    recall  f1-score   support

           0       0.83      0.82      0.83       841
           1       0.83      0.84      0.83       852

    accuracy                           0.83      1693
   macro avg       0.83      0.83      0.83      1693
weighted avg       0.83      0.83      0.83      1693
```

After Outlier Removal:



```
>dtree 0.808 (0.014)
>gb 0.749 (0.016)
>rf 0.824 (0.012)
>lr 0.703 (0.017)
>Stacked 0.832 (0.013)
```



```
              precision    recall  f1-score   support

           0       0.85      0.82      0.83       829
           1       0.83      0.86      0.85       863

    accuracy                           0.84      1692
   macro avg       0.84      0.84      0.84      1692
weighted avg       0.84      0.84      0.84      1692
```

Algorithm Comparison:

Stacking the classifiers showed a good accuracy and it had a similar accuracy to Random Forest Model. The Random Forest Model was used to carry out predictions. These classifiers are all powerful. For example, decision trees capture non-linear relationships, while random forests reduce overfitting. Gradient boosting machines learns from previous predictions, while logistic regression is a simple model in the stacking.

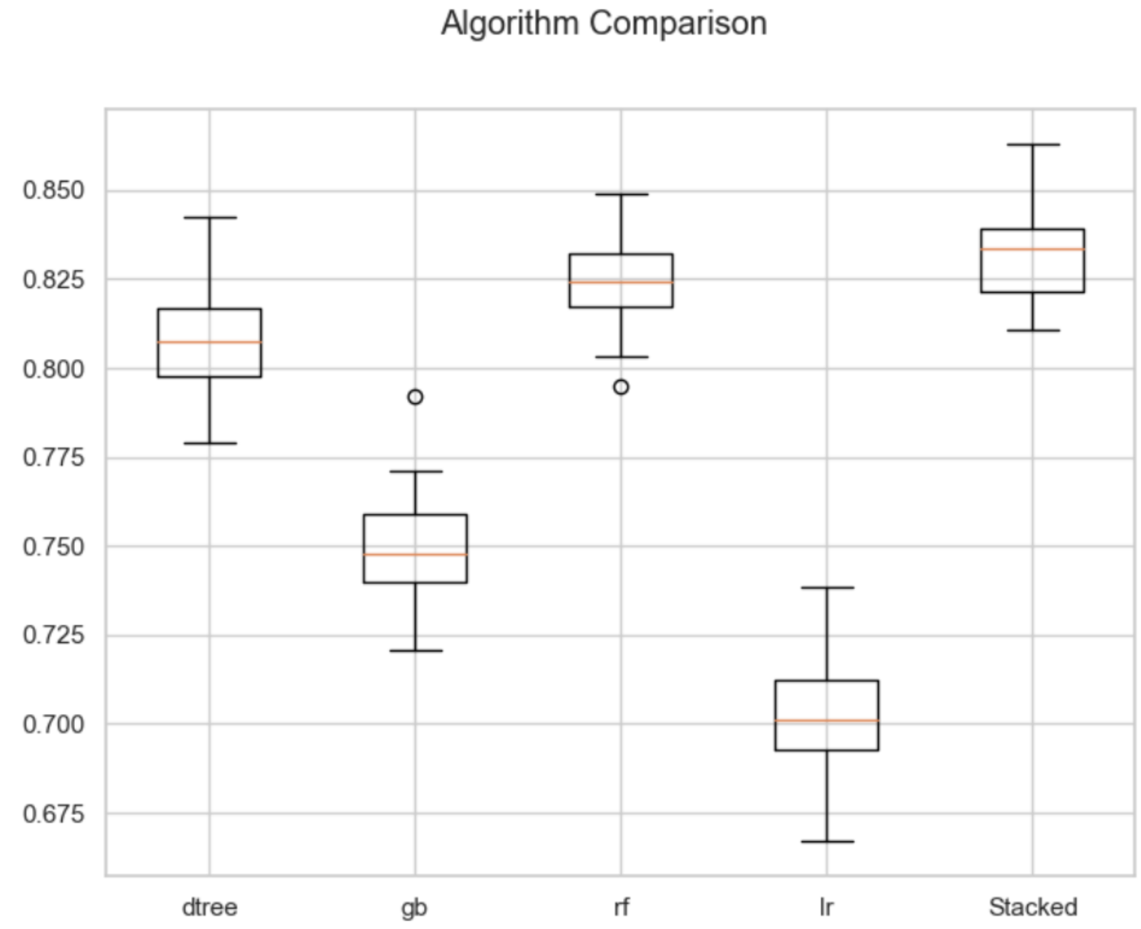


Figure 23: Algorithm Comparison

Performance Metrics and ROC Curve:

The precision of the model is 0.85, which means that of the accidents that the model classified as fatal, 85% were fatal which is a good percentage. The recall of the model is 0.86, which means that it correctly classified 86% of the fatal accidents. The f1-score of the model is 0.85, which is a good score. The model achieved an accuracy of 0.84, meaning that it accurately classified 84% of the fatal accidents.

Also, 0.837 accuracy was achieved on the validation data and 0.84 accuracy was achieved on the test data. This is an insignificant difference suggesting that the model is not overfitting and well generalized to unseen data.

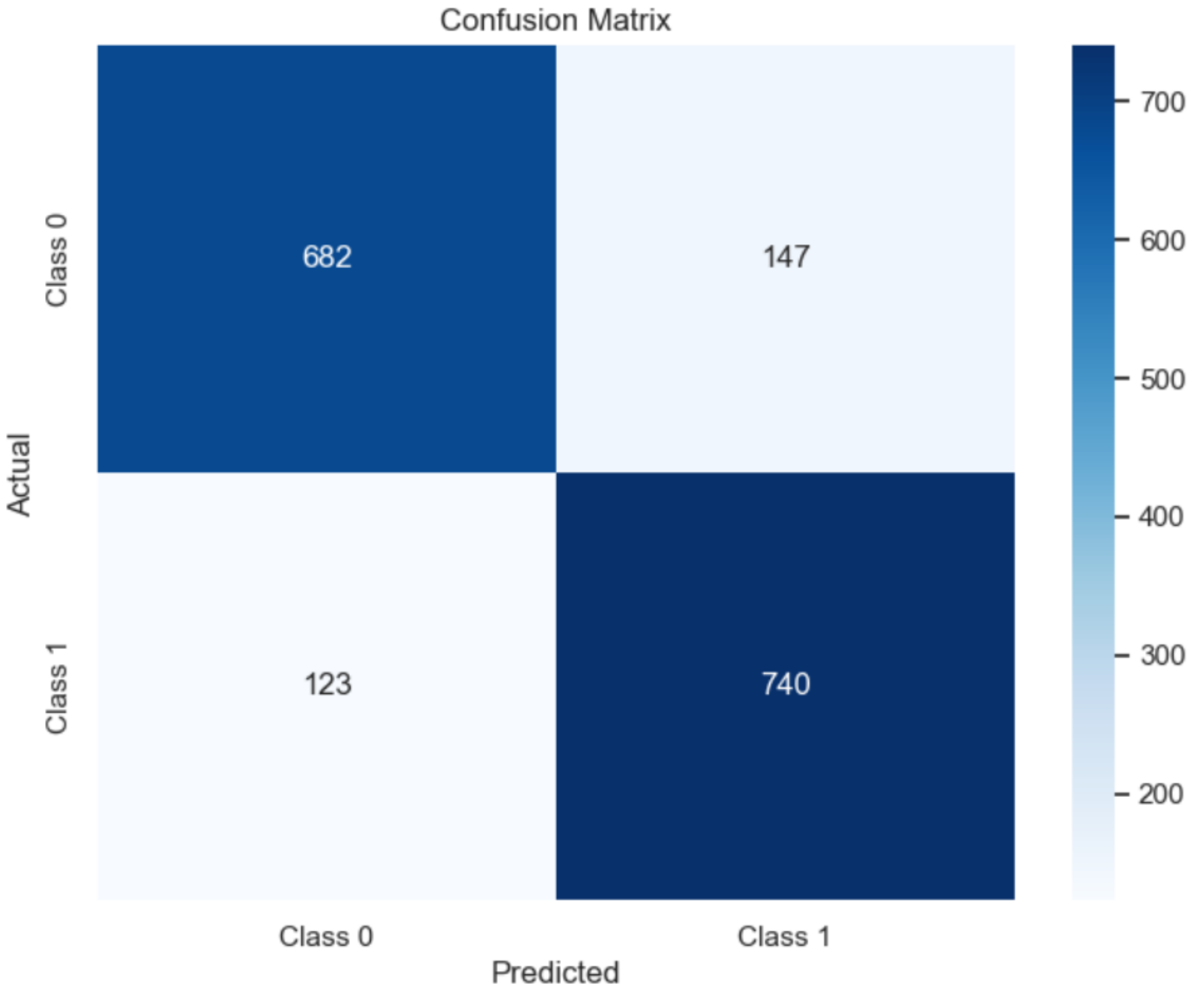


Figure 24: Confusion Matrix

The ROC curve which is a trade-off between true positive rate and false positive rate has a score of 91% which means that the model can distinguish between fatal and non-fatal accidents with a high accuracy.

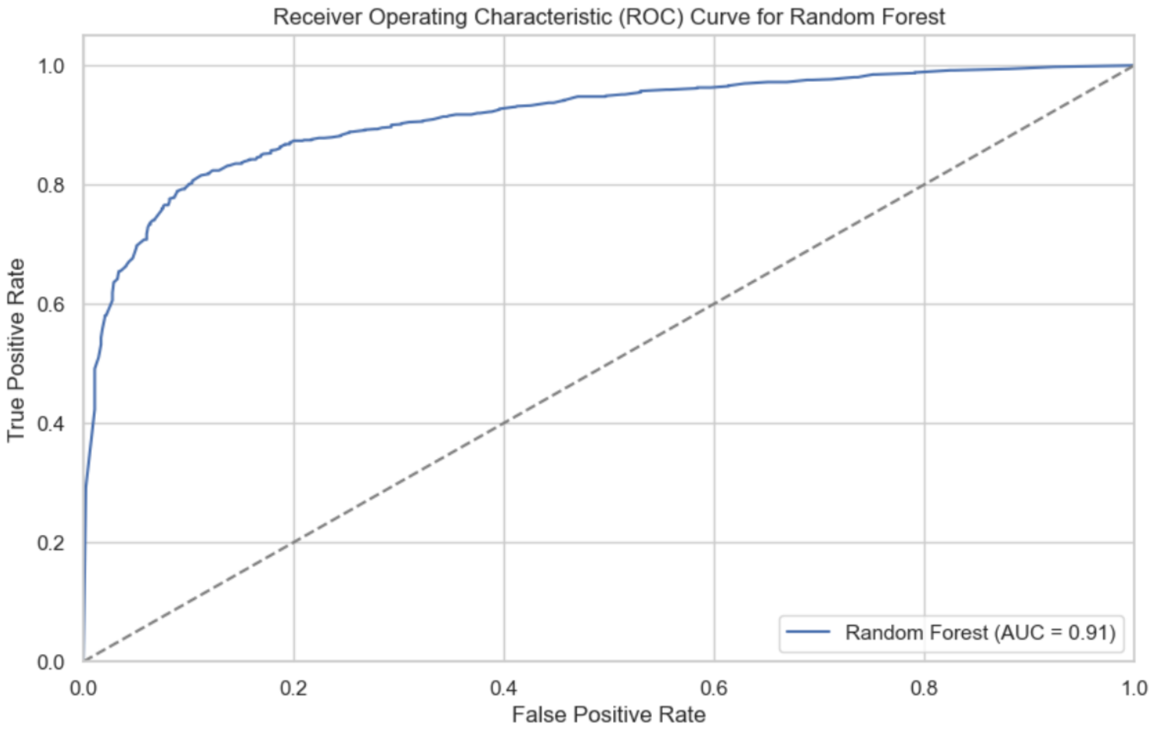


Figure 25: ROC Curve for Model

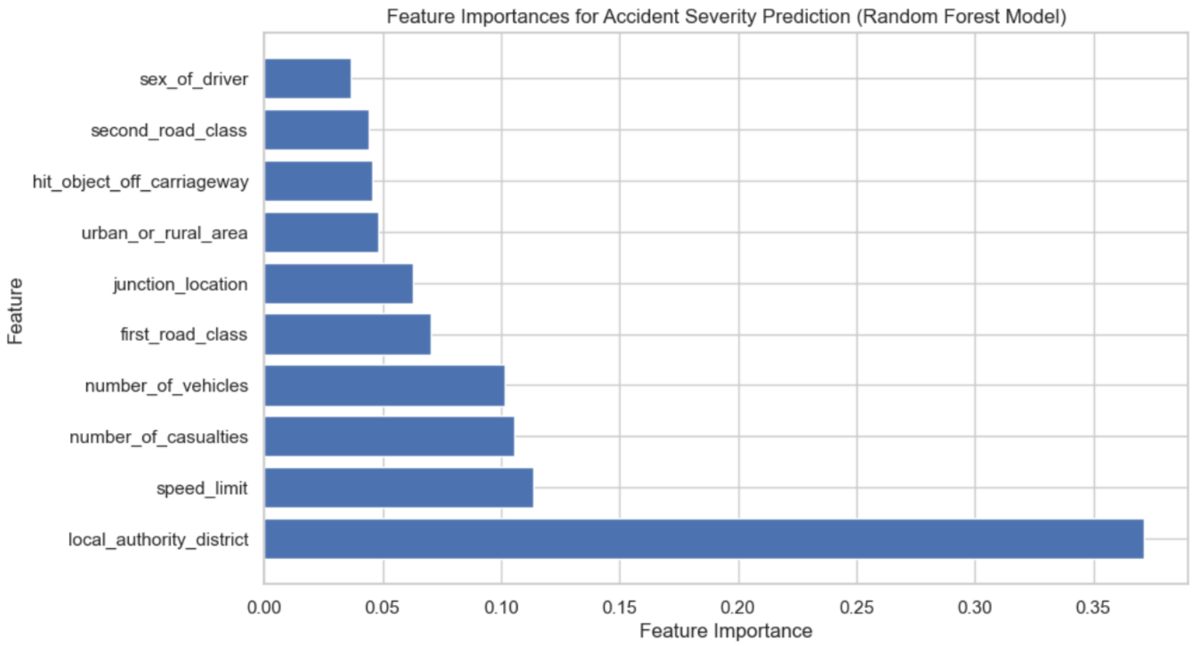


Figure 26: Feature Importance After Prediction

8. Recommendations:

- Automated Personalized Safety Recommendations based on analysis of a driver's individual risk factors should be implemented. For example, safety pop-ups on smart watches for a driver who gets drunk during weekend social activities reminding him to be less sober before driving.

- For winter periods, Sensors should be used to monitor road conditions in real time and to identify areas that are icy or slippery. Warnings can then be sent to drivers in form of notifications on how to drive on such areas.

- AI-powered cameras and sensors can be used to monitor driver behavior for signs of distracted driving, speeding, or other risky behavior.

- Elderly drivers should be made to submit their eye and blood pressure test every three years when they renew their driver’s license to be sure they are fit to keep driving.

- Crash-Maps should be used to track trends in accident occurrence over time. This information can be used to identify areas where safety interventions are needed.

References:

Cars-Data.com (No Date) Ford Fiesta Engine Capacity. Available Online: Ford Fiesta Engine capacity (cars-data.com) [Accessed 14/08/2023]

Enginert.com (No Date) Vauxhall Vectra Engine Size in Quick to Read Charts. Available Online: [1995 - 2008] Vauxhall Vectra Engine Sizes – Cool Informative Charts | Enginert [Accessed 14/08/2023]

Gov.uk (No Date) Driving Lessons and Learning to Drive. Available Online: Driving lessons and learning to drive: Overview - GOV.UK (www.gov.uk) [Accessed 14/08/2023]

Gov.uk (No Date) Historic (classic) Vehicles: MOT and Vehicle Tax. Available Online: Historic (classic) vehicles: MOT and vehicle tax: Eligibility - GOV.UK (www.gov.uk) [Accessed 14/08/2023]

Gov.uk (No Date) Renew your Driver's License if You are 70 and over. Available Online: Renew your driving licence if you're 70 or over - GOV.UK (www.gov.uk) [Accessed 14/08/2023]

Jang, K. *et al.* (2013) 'Evaluation of pedestrian safety', *Transportation Research Record: Journal of the Transportation Research Board*, 2393(1), pp. 104–116. doi:10.3141/2393-12.

Langley, J. *et al.* (2000) 'Motorcycle engine size and risk of moderate to fatal injury from a motorcycle crash', *Accident Analysis &amp; Prevention*, 32(5), pp. 659–663. doi:10.1016/s0001-4575(99)00101-3.