Big Data And Data Mining

Project Report

Introduction

This study embarks on an in-depth analysis of the UK's road accident data, with a goal to improve national road safety. By thoroughly investigating different facets of road incidents – from motorbike accidents to pedestrian involvement and the factors that affect accident severity – I'm working to unravel the hidden forces behind these often-devastating events.

My goal isn't just to enhance the existing safety measures, but to foresee and prevent fatal incidents, ultimately contributing to safer roads for everyone. I'm utilizing advanced analysis methods like association rule mining , clustering and classification model to achieve this. When all's said and done, I'll distill my findings into clear, actionable advice, offering our government new and effective ways to enhance road safety.

Analysis

Question 1

According to my analysis, Friday sees the highest frequency of accidents, as shown in Fig1.1, This trend could be attributed to the onset of the weekend, with a rise in social events, such as happy hours or gatherings that may involve alcohol, resulting in heavier traffic. Investigating the accident count by hour revealed that 5pm has the most incidents, possibly linked to rush hour at the workday's end . There's also a notable increase around 8am, representing a steep rise compared to surrounding hours, which might correspond to morning commuters.

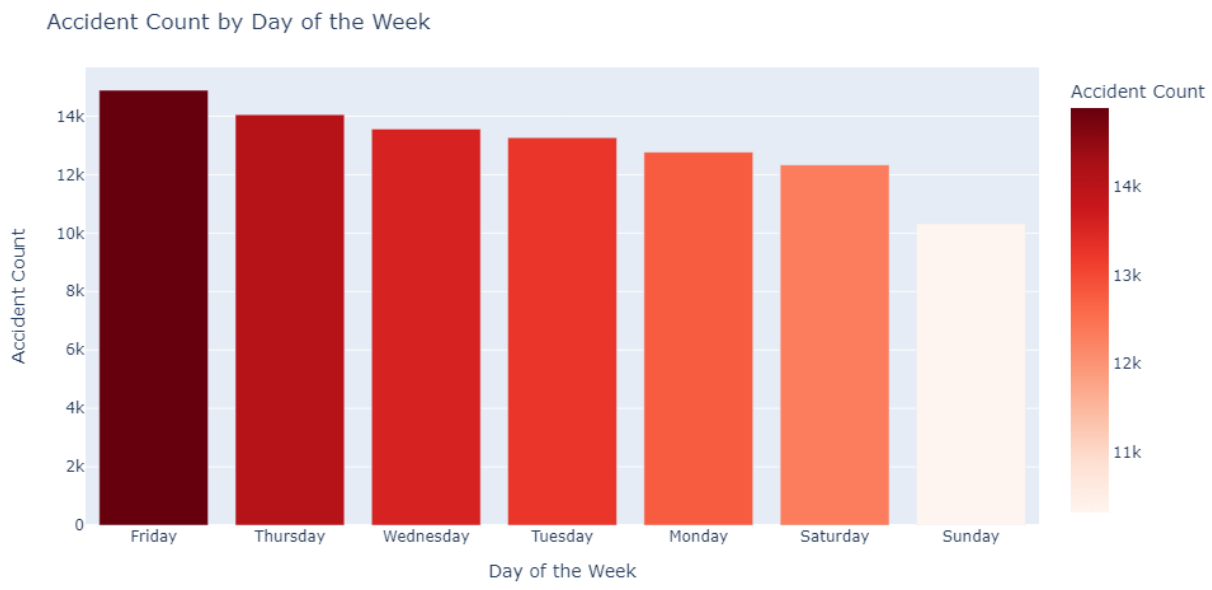

Fig1.1 Image overview of the accident count by day of the week.

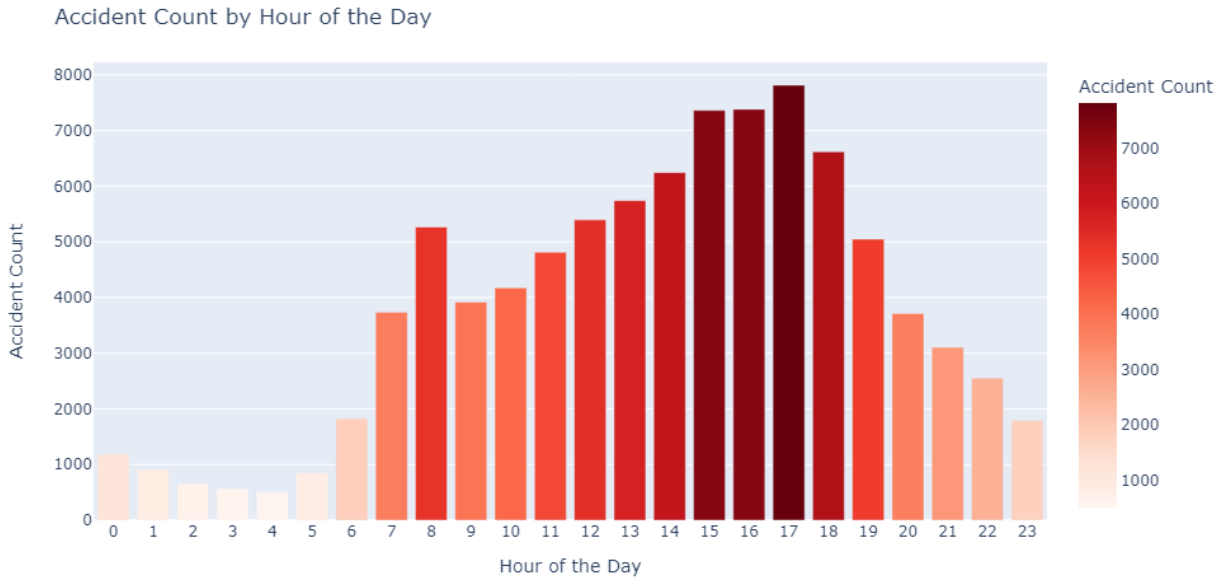

Fig1.2 Image overview of the accident count by hour of the day

Question 2

For motorcycle, an examination of the data reveals that Friday has the highest occurrence of accidents within the week, as expected. Among the various categories, motorcycles with a capacity of 125cc and under are most frequently involved. Their commonality likely reflects their affordability and accessibility, making them more attractive and economical for beginners and commuters, thus explaining their ubiquity. They are followed by motorcycles over 500cc, associated with higher speeds, and those between 125cc and 500cc, as illustrated in Fig 1.2. Moreover, a detailed analysis of the hourly distribution indicates that 5pm emerges as the most significant hour for accidents, potentially correlating with rush hour congestion, as depicted in Fig 1.3.

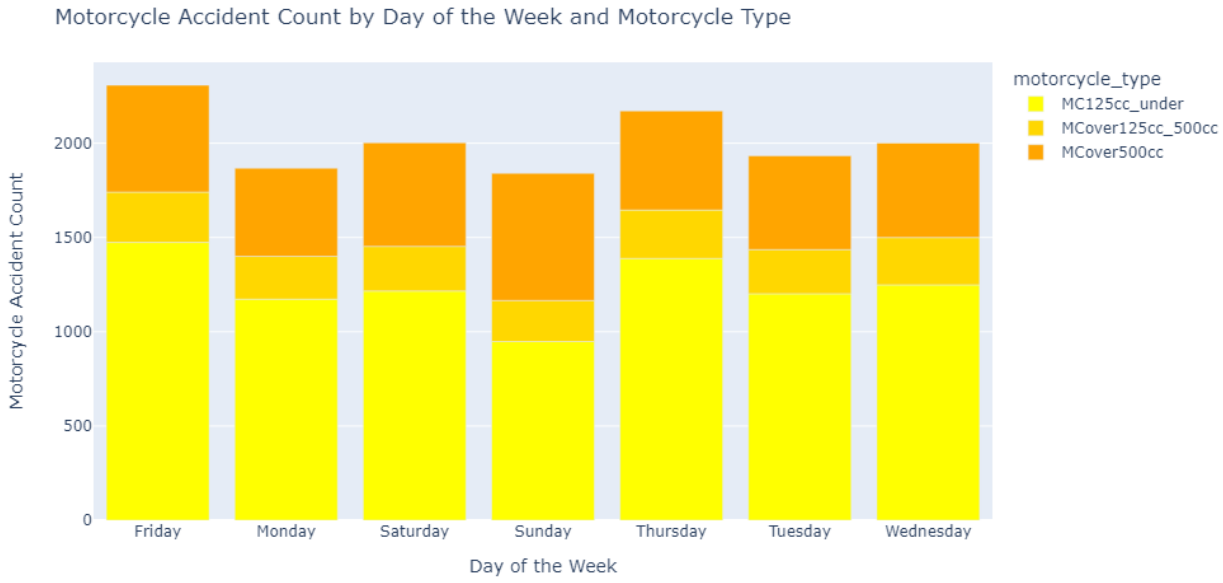

Fig1.3 Image overview of the Motorcycle accident count day of the week

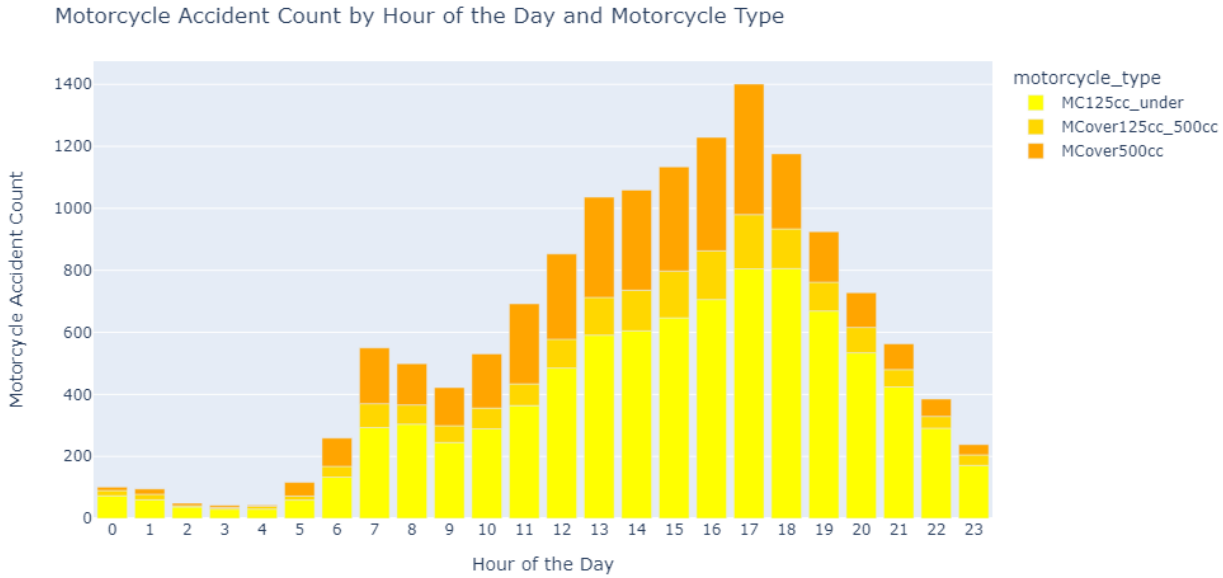

Fig1.4 Image overview of the Motorcycle accident count by hour of the day

Question 3

The observed phenomenon of increased pedestrian accidents on Fridays at 3pm, as documented in Fig 1.6, can be understood through a multifaceted analysis. School dismissal at this hour often leads to a surge of young pedestrians who may be less experienced in navigating traffic safely. This is exacerbated by an influx of vehicles as parents and guardians arrive to pick up children, creating a more congested and potentially hazardous traffic environment.

Simultaneously, Fridays may mark a unique pattern in workplace behavior. Some companies encourage or allow employees to leave early, resulting in unusual traffic patterns and greater pedestrian activity. This change in routine can cause congestion, leading to heightened risk for accidents.

Additionally, public transportation may see higher usage at these times, further contributing to increased pedestrian movement. More people navigating streets to reach bus stops or train stations add to the complexity of the traffic landscape.

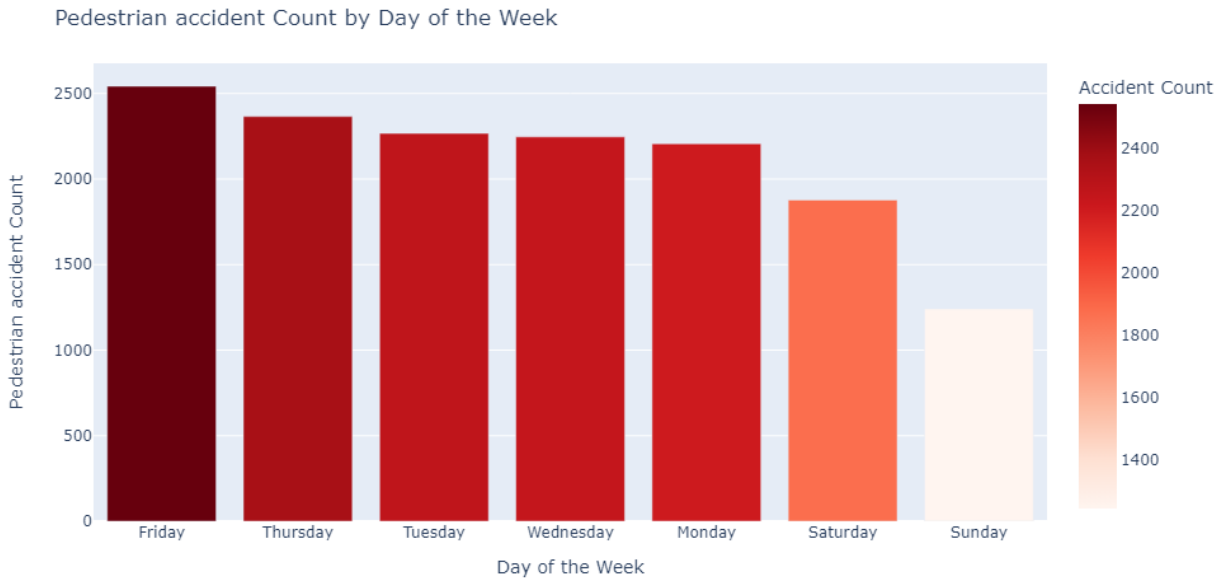

Fig1.5 Image overview of the Pedestrian accident day of the week

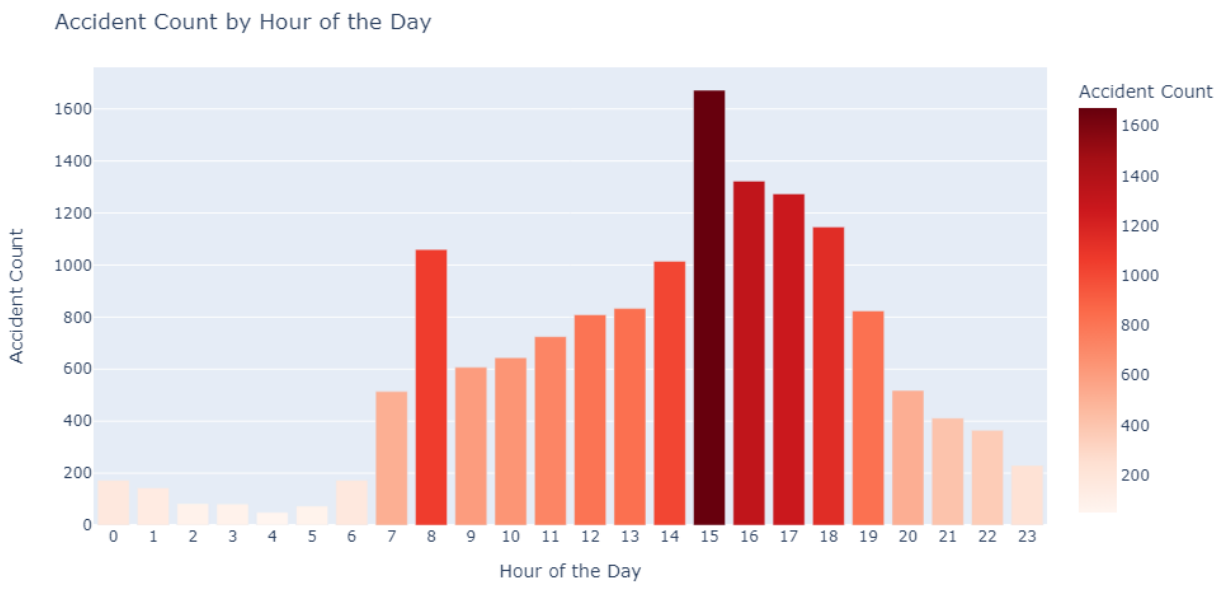

Fig1.6 Image overview of the Pedestrian accident count by hour of the day

Apriori

Question 4

The Apriori algorithm was utilized in this study to explore the impact of selected variables on accident severity. Drawing from accident table, key factors such as road type, speed limit, road surface conditions, weather conditions, sex of the driver, and accident severity were analyzed.

Care was taken to address outliers and missing values, manually identified within some of these columns. An inherent correlation between road surface conditions and speed limit guided the replacement of -1 values, with the median speed limit within road surface condition groups used as a substitute. Similarly, the correlation between weather conditions and road surface condition informed the cleaning of the road surface conditions column. Lastly, 13 rows with undefined sex of the drivers were excluded from the analysis.

After cleaning the data, I applied one-hot encoding to the chosen features and created a new dataframe called 'encoded_AccVehCas_data2020.' I then used the Apriori algorithm on this dataframe, setting a 20% minimum support threshold to find common patterns or combinations of features. Next, I used association rules to uncover how these features relate to one another, particularly focusing on their antecedents and consequents, and using lift as the metric to measure these relationships.

The resulting dataframe, was filtered to examine the links between specific combinations of features (antecedents) and types of accident severity (consequents). The table below provides details of some of these associations:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	a
3	(sex_of_driver_2.0)	(severity_3.0)	0.249862	0.795477	0.206397	0.826041	1.038422	0.007637	1.175695	0.049325	
0	(speed_limit_30.0)	(severity_3.0)	0.569821	0.795477	0.465018	0.816078	1.025897	0.011738	1.112005	0.058680	
20	(speed_limit_30.0, road_type_6.0)	(severity_3.0)	0.456875	0.795477	0.368028	0.805533	1.012640	0.004594	1.051706	0.022983	

Fig1.7 Image overview of the Pedestrian accident count by hour of the day

Antecedent: (sex_of_driver_2.0) and Consequent: (severity_3.0)

Lift: 1.038422

When the sex of driver is recorded as 2, there's a 3.84% higher likelihood that the accident severity is compared to if these two factors were independent. This suggests a minor but noticeable association between the driver's sex and the severity of the accident.

Antecedent: (speed_limit_30.0) and Consequent: (severity_3.0)

Lift: 1.025897

Accidents that occur with a 30 speed limit are 2.59% more likely to result in severity level 3.0 compared to if speed limit and severity were unrelated. This might imply that accidents in areas with this speed limit tend to be of a slight severity.

Antecedent: (speed_limit_30.0, road_type_6.0) and Consequent: (severity_3.0)

Lift: 1.012640

When the speed limit is 30 and the road type is 6.0, there is a 1.26% increased likelihood that the accident severity will be classified as 3, compared to if these factors were independent. This relationship hints at a nuanced interaction between speed limit and road type, impacting the severity of accidents.

Clustering

Question 5

In order to identify accidents within our region, encompassing Kingston upon Hull, Humberside, and the East Riding of Yorkshire, I began by fetching the relevant data through filtering based on police regions. Then, I employed the Elbow Method, a technique used to find the optimal number of clusters. This method pinpointed five as the optimal number of clusters to capture the patterns in our data, as documented in Fig 1.8.

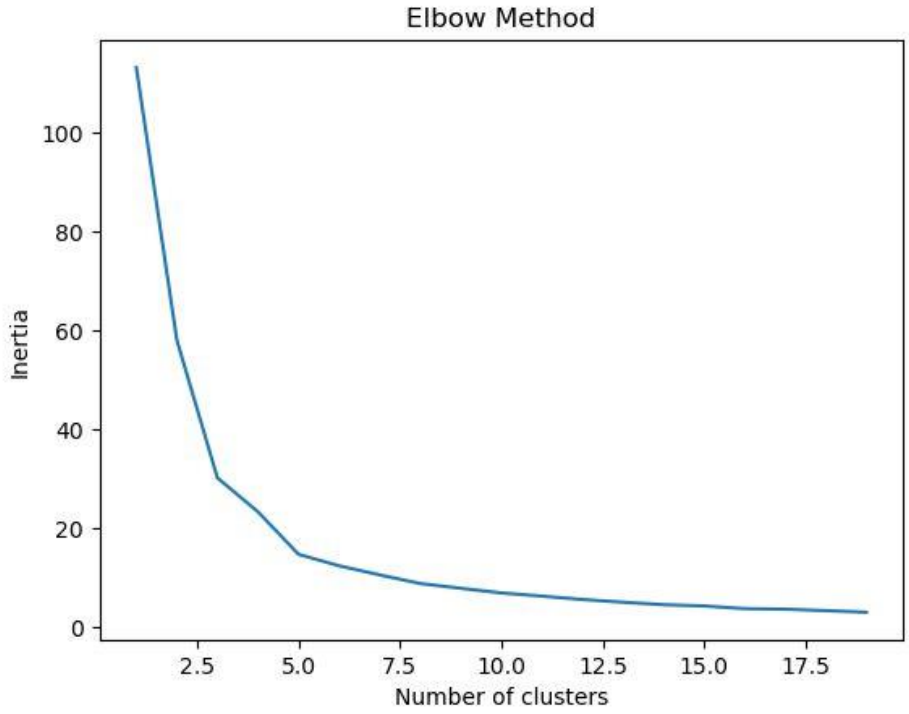

Fig1.8 Image overview of the Elbow Method

With this information in hand, I applied the K-means clustering algorithm, enabling a detailed examination of the distribution of accidents within the defined region. The analysis identified specific patterns in our region's accident distribution, as shown in Fig 1.9. Kingston upon Hull shows a high density of accidents, likely due to its location near the city center and major roads leading to Hull. Similarly, high accident density is found in Scunthorpe and North East Lincolnshire.

The fourth cluster reveals sparse accidents between Bridlington and Driffield, while the fifth cluster is scattered around Pocklington and Goole. These patterns might reflect variations in traffic congestion, road layout, and road quality.

The performance of the k-means clustering was evaluated using 3 internal measures:

The Davies-Bouldin Index of 0.627 indicates a moderate separation between clusters,

The Silhouette Coefficient of 0.622 suggests that the clusters are reasonably well defined and separated

The Calinski Harabasz Score of 2757.194 further confirms that the clusters are dense and well-separated, reinforcing the validity of the clustering solution.

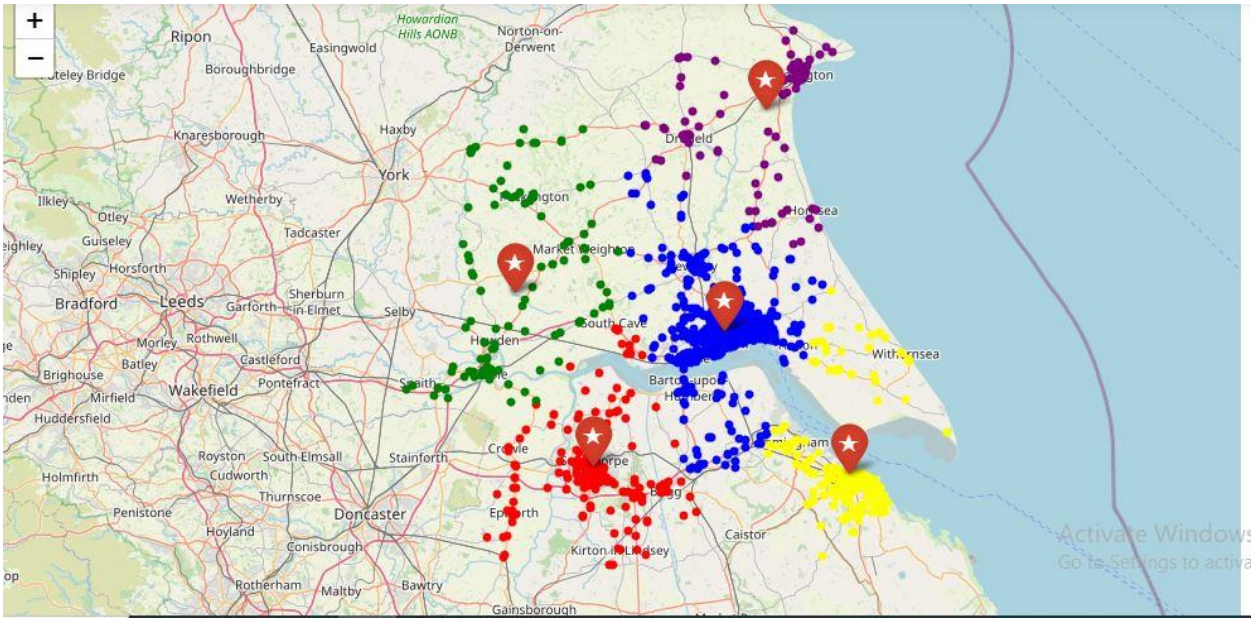

Fig1.9 Image overview of K-Means Clustering

Outliers

Question 6

Later in this study, i will utilize the accident table from the accident database to create a classification model that can accurately predict severe injuries resulting from road traffic accident. It's crucial to clean and preprocess the data, ensuring its quality and reliability. A key aspect of this preprocessing is the detection and handling of outliers. To detect outliers, I used Isolation Forest, After employing the Isolation Forest method to detect outliers, i found that the method, despite its strengths, was not entirely comprehensive. As evidenced in Figure 1.10 and 1.11, there remained undetected outliers even after adjusting the contamination level to 30%. In light of this, we leveraged domain knowledge to manually pinpoint the remaining anomalies.

```
12 columns detected as outliers

['local_authority_district', 'speed_limit', 'junction_control', 'pedestrian_crossing_human_control', 'pedestrian_crossing_physi
cal_facilities', 'light_conditions', 'weather_conditions', 'road_surface_conditions', 'special_conditions_at_site', 'carriagewa
y_hazards', 'did_police_officer_attend_scene_of_accident', 'iforest_outlier']
```


Fig1.10 Image showing the total number of column where outliers detected by outliers

```
local_authority_district
speed_limit
junction_detail
junction_control
pedestrian_crossing_human_control
pedestrian_crossing_physical_facilities
light_conditions
weather_conditions
road_surface_conditions
special_conditions_at_site
carriageway_hazards
did_police_officer_attend_scene_of_accident
iforest_outlier
13  columns actually have outliers
```

Fig1.11 Image showing the total number of column where outliers were manually detected

Our data cleaning process unfolded in a systematic, column-by-column manner:

1. **Local_Authority_District**: Upon examination, It was found that unique local_district_authority areas align with each police_code. Therefore, the dataframe was grouped by police_code, medians were computed, and -1 values in local_district_authority within these groups were replaced with there medians.
2. **Speed_Limit**: Recognizing the intrinsic correlation between road surface conditions and speed limits, -1 values in the 'speed_limit' column were handled adeptly. The median speed limit corresponding to each road surface condition was used as a substitute, guaranteeing accurate imputation.
3. **Junction_Control**: In our dataset, over 38,000 rows had a value of -1. As outlined by the Department of Transport (2011), if "Junction Detail" is coded as 00, "Junction_Control" should be left blank. However, in our dataset, 'junction_control' was labeled as -1 when 'junction_detail' had a 00 code. Since some machine learning model may not be able to process blanks, we coded 'junction_control' with a unique value of 0 whenever 'junction_detail' was 0.
4. **Road Surface Condition**: Cleaning of this column was influenced by the observed correlation between weather conditions and road surface states.
5. **Other Columns**: For the remaining columns with outliers, particularly those with a value of -1, the strategy was to replace them with the median value of the respective column.

Classification Model

Question 7

After preprocessing the 'accident' table, the target variable was reclassified from three categories to just 'fatal' and 'non-fatal'. I employed features selection Model SelectKBest, to pinpoint the most crucial features from the dataset, as depicted in Fig 1.12.

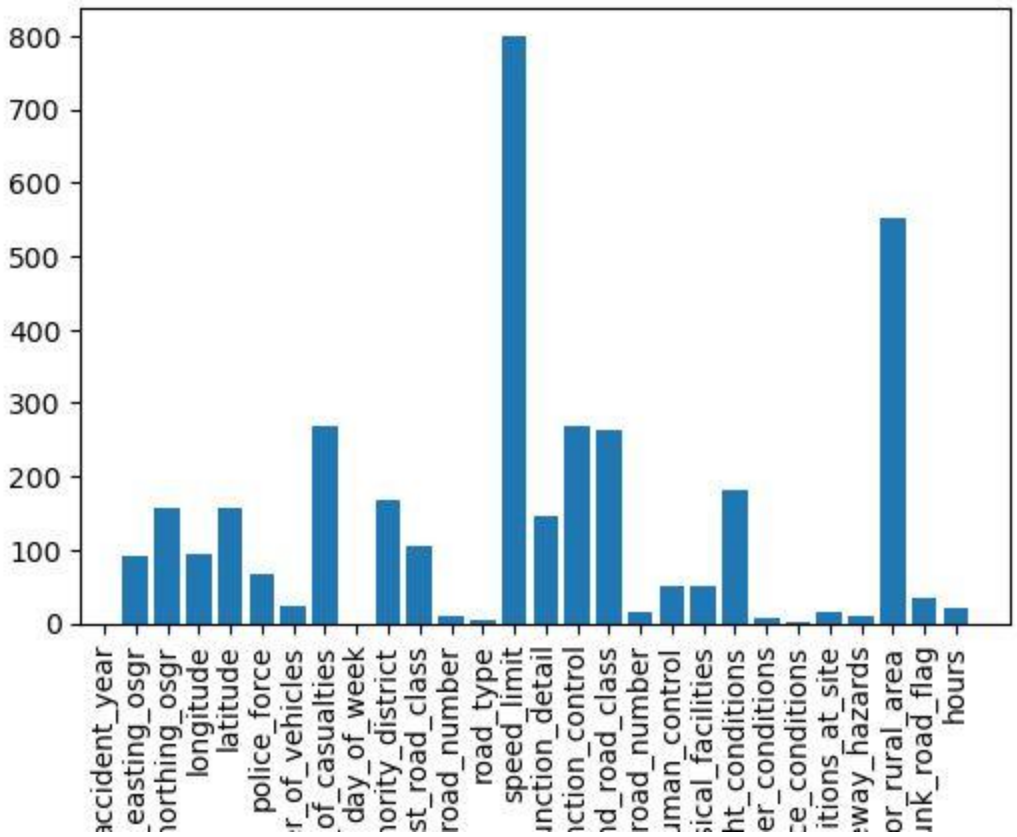

Fig1.12 Image showing the features and their relative importance

Analysis of the target variable's distribution revealed a significant imbalance—'non-fatal' instances are notably fewer than 'fatal' ones. As illustrated in Fig 1.13.

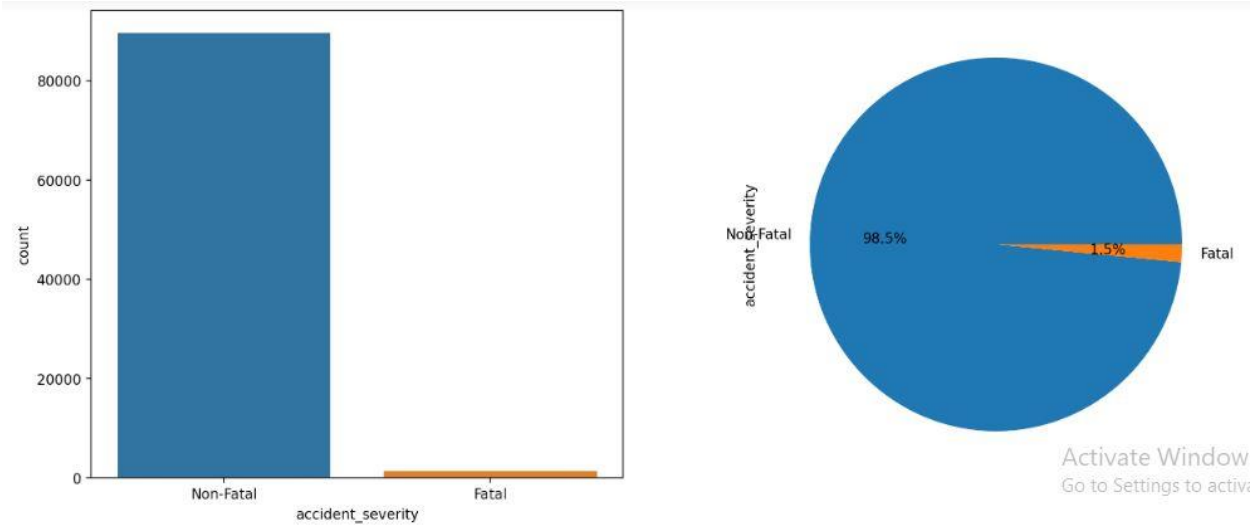

Fig1.13 Image showing imbalanced target variables

This data imbalances can compromise my model's efficacy, particularly in discerning features of the underrepresented class which in this case is Fatal, which often results in misclassifications. This adversely affects the model's overall predictive accuracy. To address this challenge, I employed the RandomUnderSampler technique. This method strategically reduces instances from the majority class while maintaining authenticity and integrity, to achieve dataset equilibrium, enhancing model robustness. The is detailed in Fig 1.13.

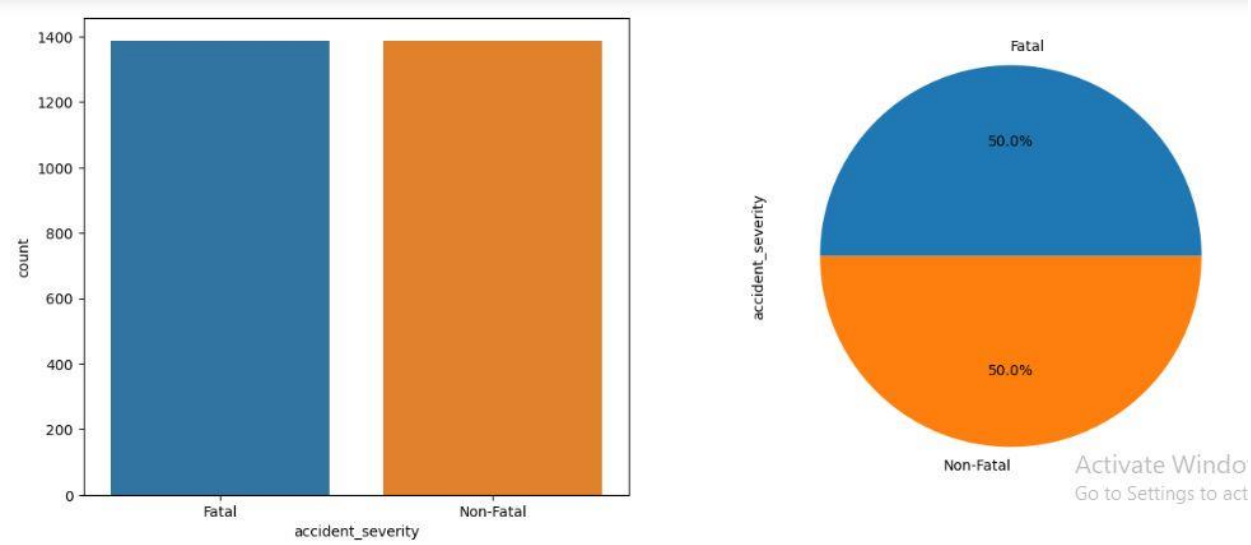

Fig1.13 Image showing balanced target variables after Random Sampler

In the early stage of building best model, the primary model selected was a decision tree. This choice was based on its adaptability and ease of interpretation. Regrettably, despite our efforts, it attained a modest 60% accuracy. This level of performance, which can be categorized as average, is detailed and broken down in the table provided below.

	precision	recall	f1-score	support
Fatal	0.62	0.58	0.60	284
Non-Fatal	0.59	0.62	0.60	272
accuracy			0.60	556
Macro avg	0.60	0.60	0.60	556
Weighted avg	0.60	0.60	0.60	556

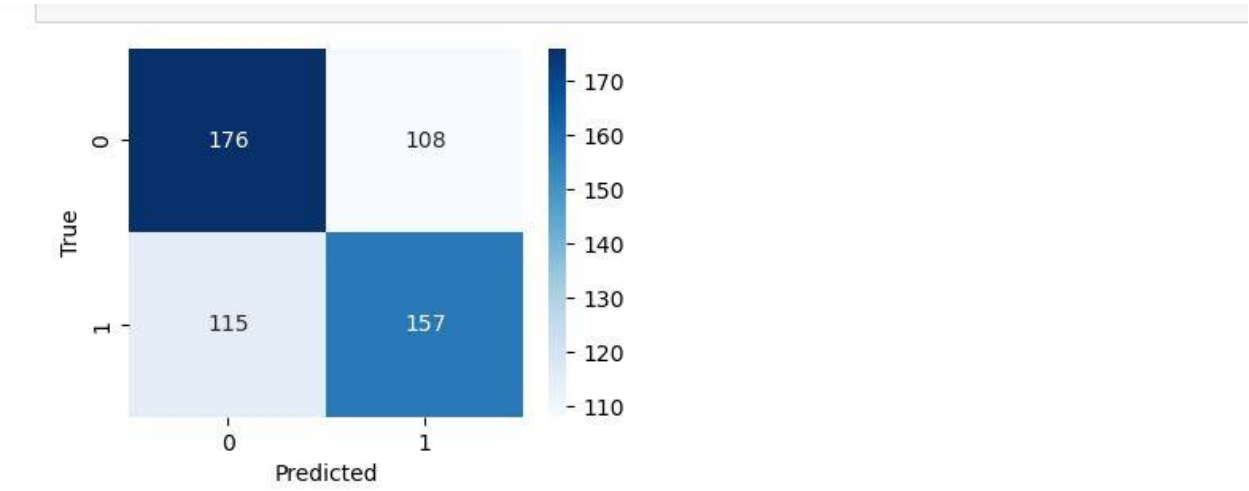

Fig1.14 Image showing confusion matrix

Furthermore, in my continuous effort to explore avenues to enhance classification accuracy, i turned to the Voting Classifier. This ensemble learning technique combines predictions from diverse models to improve the overall performance. The ensemble integrated the Decision Tree, KNN, GradientBoostingClassifier and Random Forest models, employing a cross-validation set at 5 folds. As anticipated, this approach led to an improved average accuracy of 69%, manifesting a robust classification performance, as illustrated in the table and a higher confusion matrix in fig1.15 below.

	precision	recall	f1-score	support
Fatal	0.67	0.77	0.72	284
Non-Fatal	0.72	0.60	0.65	272
accuracy			0.69	556
Macro avg	0.69	0.69	0.68	556
Weighted avg	0.69	0.69	0.68	556

1. **Precision (0.67)**: 67% of the model's fatal predictions were accurate, pointing to some overestimation.
2. **Recall (0.77)**: The model correctly identified 77% of all true fatalities, missing 23%.
3. **F1-Score (0.72)**: The F1 score, a crucial balance metric, reflects a commendable harmony between precision and recall at 0.72. It indicates the model's ability to manage false alarms while capturing genuine cases.

Given these metrics, our model showcases a fairly good performance in predicting fatal accidents

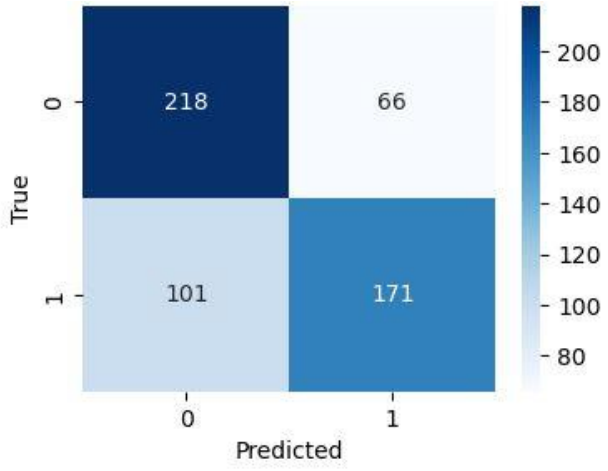

Fig1.15 Image showing confusion matrix after using ensemble learning

Recommendation

1. Adopting Advanced Traffic Management Systems that can adapt to changing traffic conditions and optimize signal timings in areas with high accident density shown in the clustering such as Kingston Upon Hull, Scunthope and North East Licolnshire.
2. In high accident density area such as city centers, collaboration with local businesses to stagger work hours, thus reducing the rush during peak times
3. Given that Fridays have the most accidents, law enforcement organisations could enhance patrols or visibility on this day to discourage risky driving and offer prompt accident response.
4. Given that most accidents involve motorcycles under 125cc, an awareness campaign for new and younger riders can be valuable. Consider offering safe-riding courses or incentives for safety training in this segment.
5. Given the peak in pedestrian accidents at 3pm, likely aligned with school closing times, it's crucial to introduce a concise road safety curriculum in schools. This program should emphasize on Safe pedestrian practices, The significance of using designated crossings, Dangers of impulsively entering the roadway.

Reference

Department of Transport, 2011. *STATS19 forms and guidance*. Available at: https://www.gov.uk/government/publications/stats19-forms-and-guidance [8/07/2023].