ROAD TRAFFIC ACCIDENT REPORT

INTRODUCTION

This report delves into the comprehensive analysis of 2020 road traffic accident data in the UK with the aim of advising government agencies on road safety improvements and developing a predictive model for accident fatality. Through the utilization of data analysis and machine learning techniques, the report provides insights into significant hours and days of the week when accidents occur. It further explores related patterns to motorbike accidents, pedestrian involvement, the impact of selected variables on accident severity, outliers and clusters. The analysis ultimately informs the creation of a predictive model for fatal injuries sustained in road accidents.

METHODOLOGY & ANALYSIS

Data Collection & Pre-processing

The foundation of the analysis is from the UK accident data SQLite database containing comprehensive records of road traffic accidents in Great Britain. The Data extraction focused on the year 2020. The road traffic accident statistics form served as a key reference, ensuring accurate interpretation of data. During pre-processing, efforts were made to handle missing values, address inconsistencies and merge relevant datasets to form a comprehensive analytical foundation.

Accidents by the Hour

The analysis of accidents by hour of the day reveals interesting patterns in the distribution of accidents throughout the 24-hour period. Figure 1 indicates that the hours between 3 PM and 6 PM experience the highest frequency of accidents. Specifically, the hour of 5 PM stands out as the peak hour, with nearly 8,000 reported accidents. On the other scale, the early morning hours, specifically 3 AM and 4 AM, exhibit the lowest rates of accidents.

Accident by the Day

When examining the number of accidents by day of the week, it becomes evident that certain days exhibit higher accident rates than others. The bar chart depicting the distribution of accidents by day (Figure 2) highlights that Thursday records the highest number of accidents, closely followed by Friday. In contrast, Sunday reports the lowest number of accidents among the days of the week.

This discrepancy suggests that certain weekdays might witness heightened traffic and driving activities, possibly due to increased work-related commuting activities. The lower number of accidents on Sundays could be attributed to reduced traffic volumes and more cautious driving behaviour during weekends.

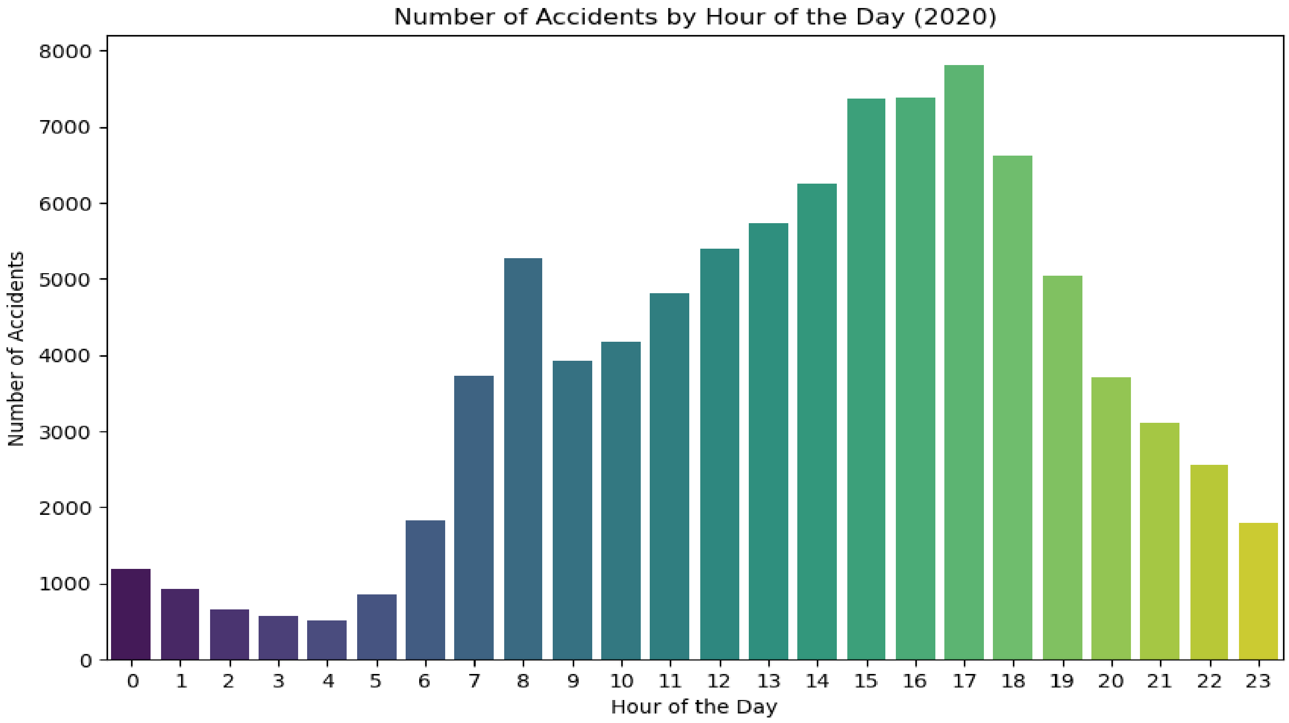


Fig 1. Number of Accidents by Hour of Day

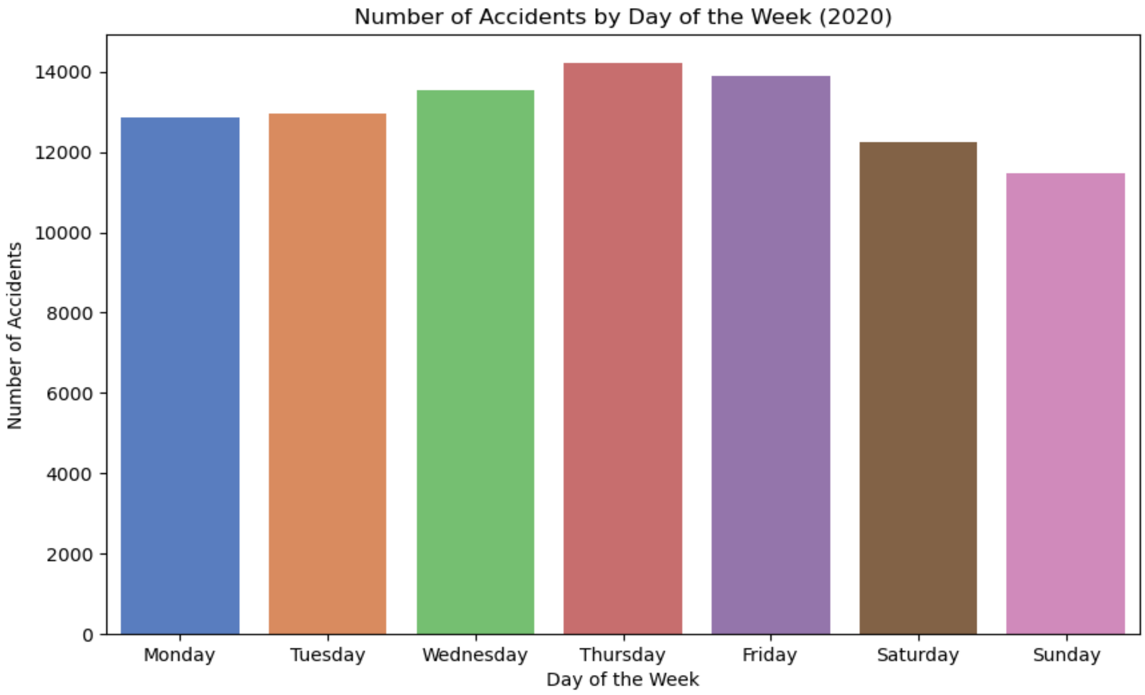


Fig 2. Number of Accidents by Day of the Week

Time of Day Influence on Motorbike Accidents

The consistent pattern observed across different categories of motorbikes points to a common trend regarding the time of day when motorbike accidents are more likely to occur. The highest occurrence of accidents around 5 PM remains prominent across all motorbike categories. This consistency reinforces the notion that certain hours of the day are associated with an increased risk of accidents regardless of the motorbike's engine capacity.

The distinct peaks between 7 and 8 AM, followed by a decrease and another surge around 10 AM, highlight the significance of morning rush hours in contributing to motorbike accidents.

Day of Week Influence on Motorbike Accidents

The pattern observed for Motorbikes with a cubic capacity between 50cc and 500cc remains consistent with most accidents occurring on weekdays with the exception of Motorbikes with over 500cc. Motorcycles with larger engine displacements, such as those over 500cc, are often associated with recreational riding and leisure activities and perhaps, alcohol consumption from social gatherings. Riders of these powerful motorcycles might choose to take longer trips or explore scenic routes on weekends, including Sundays. This increased recreational riding could lead to higher exposure to accident risks.

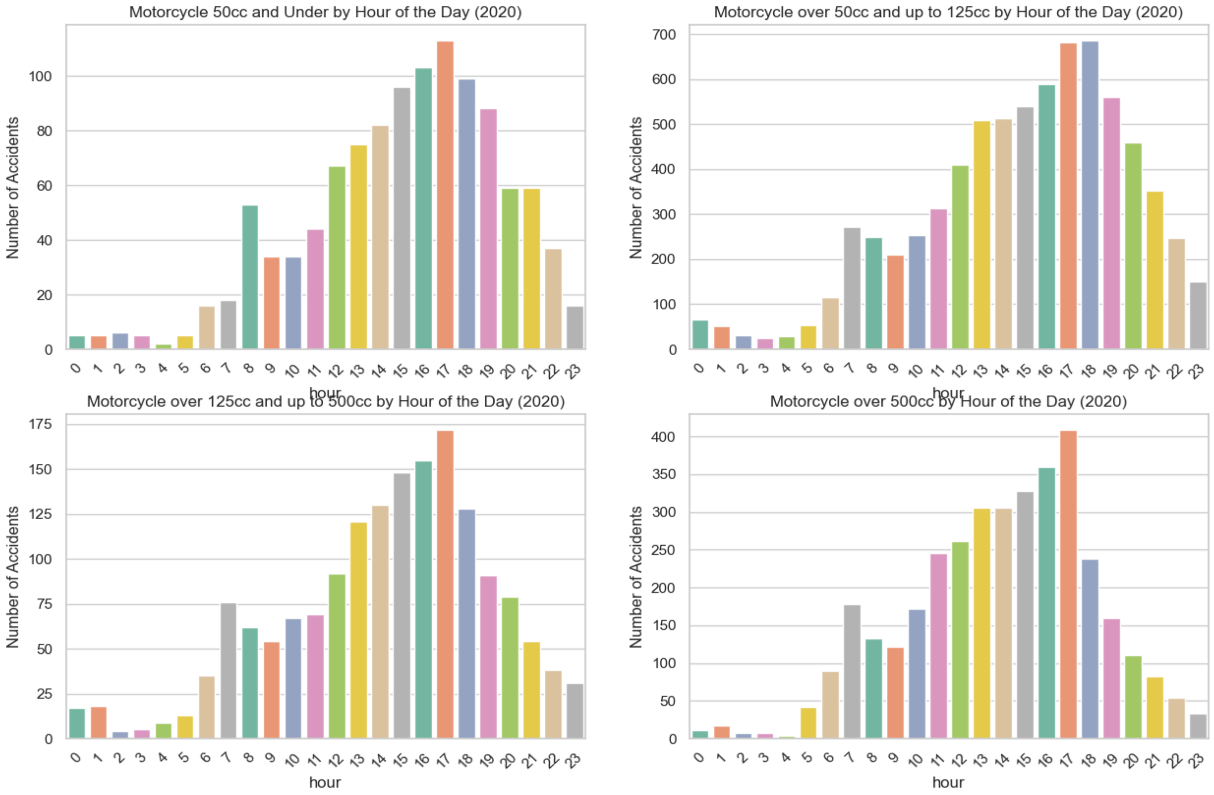
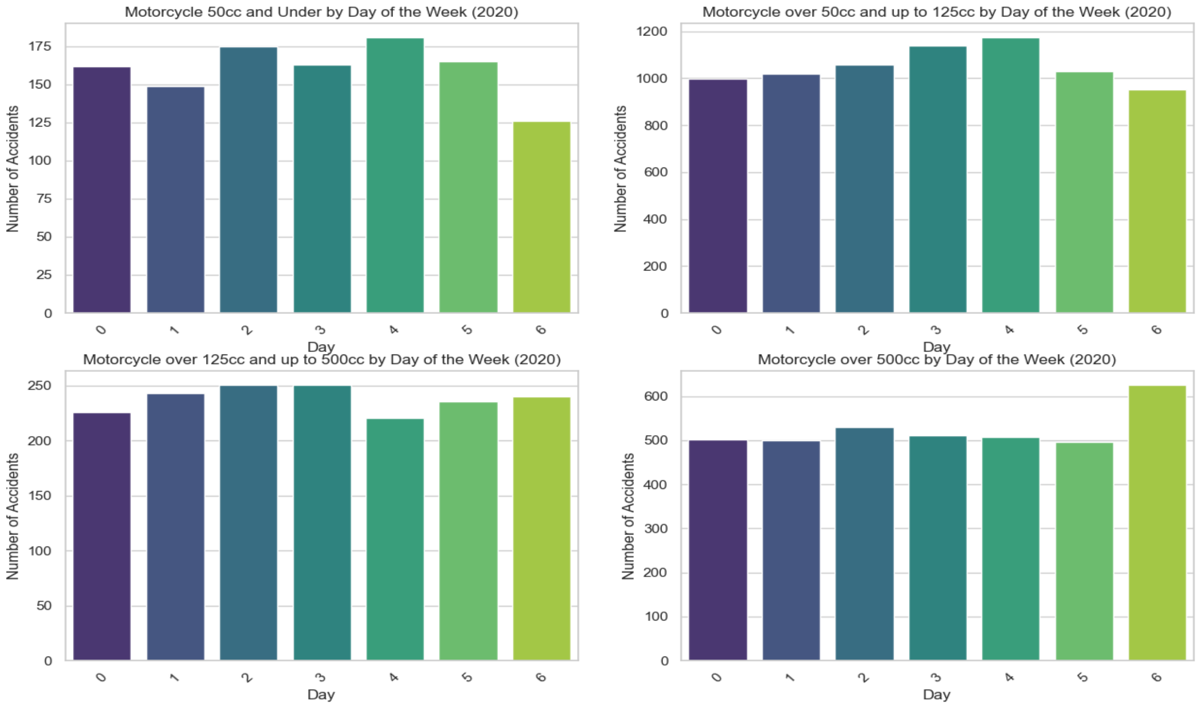


Fig 3. Motorbike Accidents by hour of day


Motorcycle 50cc and Under by Day of the Week (2020)

Motorcycle over 50cc and up to 125cc by Day of the Week (2020)

Motorcycle over 125cc and up to 500cc by Day of the Week (2020)

Motorcycle over 500cc by Day of the Week (2020)

Fig 4. Motorbike Accident by day of week

Pedestrian Accidents

The analysis of pedestrian accidents throughout the day reveals distinct patterns that hold significant implications for safety interventions:

During peak hours, notably, the span between 3 pm and 6 pm marks the zenith of pedestrian accidents. Among these hours, 3 pm emerges as a critical time with the highest accident count. This is indicative of heightened risks during school dismissal hours and the bustling afternoon traffic, warranting focused safety measures.

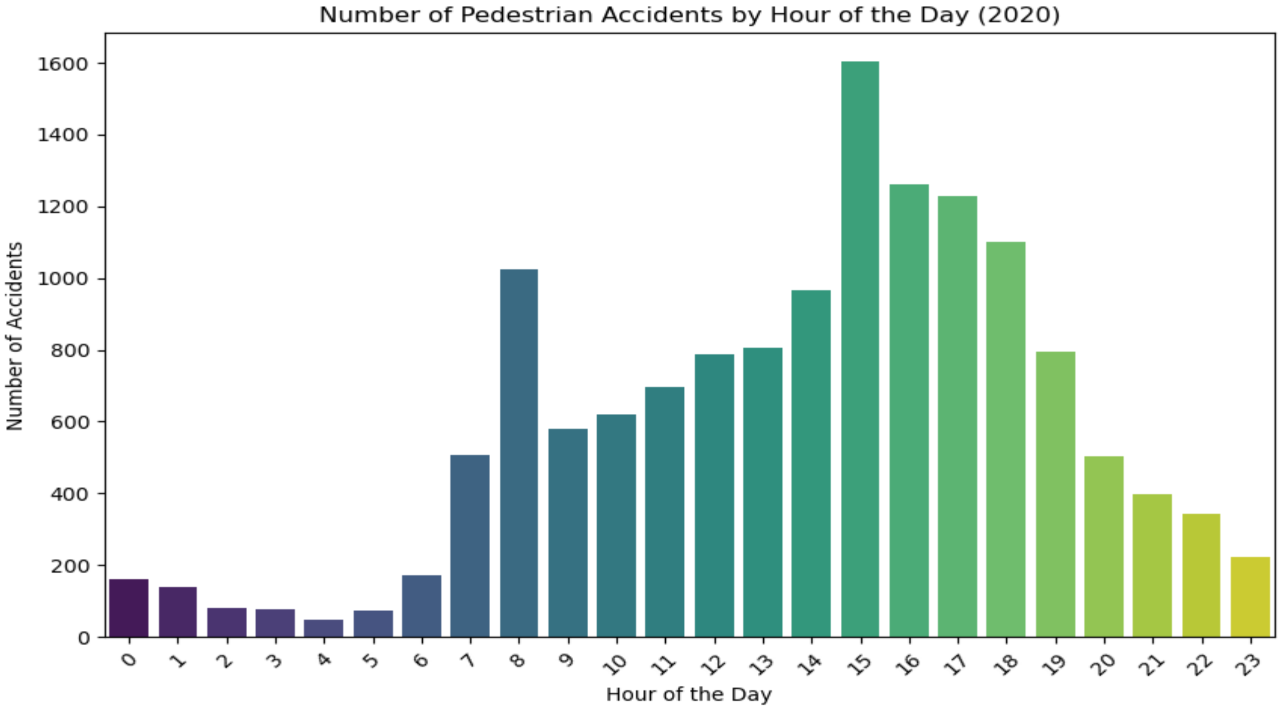


Fig 5. Pedestrian Accidents by hour of the day

Impact of Selected Variables on Accident Severity

To explore the influence of selected variables on accident severity, the Apriori algorithm was employed to uncover meaningful patterns and associations. The process involved data preparation, feature selection, and analysis of frequent itemsets and association rules.

Initially, feature selection was conducted to identify relevant attributes affecting accident severity. Considering the dataset's imbalanced nature, the data was balanced using the RandomUnderSampler from the imblearn library before feature selection. The SelectKBest method was then utilized to identify the top 10 significant features impacting accident severity. This approach facilitated the extraction of key variables that were subsequently used in the Apriori algorithm.

Severity Distribution

The dataset is characterized by varying degrees of accident severity. Specifically, severity_3 (slight accidents) has the highest support value of 0.783, followed by severity_2 (moderate accidents) with support of 0.201, and severity_1 (fatal accidents) with the lowest support of 0.015.

Certain speed limits were found to have a significant influence on accident severity. For instance, accidents occurring at speed_limit_30 exhibited a substantial support value of 0.573, suggesting a notable occurrence frequency of accidents at this speed limit. In contrast, accidents at speed_limit_20 had a lower support value of 0.123.

Association Rules Analysis

The application of the Apriori algorithm has revealed significant association rules among selected variables and different levels of accident severity, specifically fatal injuries (severity_1) and slight injuries (severity_3). These rules offer valuable insights into the relationships between variables and how they contribute to varying degrees of accident severity.

For instance, the rule "severity_1 → police_in_scence_1" indicates a strong confidence level of approximately 86.99%. This suggests that in cases of accidents resulting in fatal injuries (severity_1), there's a high likelihood that a police officer attended the scene (police_in_scence_1).

Another insightful rule is "speed_40 → severity_2," with a confidence of about 21.93%. This implies that accidents occurring at a speed limit of 40 mph (speed_40) are associated with accidents resulting in slight injuries (severity_3). Conversely, the rule "severity_2 → speed_40" suggests that when accidents result in slight injuries (severity_3), they are more likely to occur at a speed limit of 40 mph (speed_40).

Cluster Analysis of Accidents in Kingston upon Hull, Humberside, and the East Riding of Yorkshire

In the pursuit of understanding the distribution of accidents in the region encompassing Kingston upon Hull, Humberside, and the East Riding of Yorkshire, a comprehensive approach was taken involving clustering. This method is a valuable analytical tool that groups similar data points together, offering insights into patterns within the dataset.

Relevant features (longitude and latitude) were selected from the dataset and standardized using StandardScaler

Clustering Techniques Utilized

Three clustering techniques were employed: K-Means, Agglomerative Clustering, and DBSCAN. All three methods converged on similar cluster patterns

The K-Means approach, facilitated by the elbow method, initially suggested an optimal number of clusters as 3. However, upon implementation, the KMeans algorithm revealed 5 clusters. This deviation underscores the complexity of the data, as it implies subtle but significant variations beyond what the elbow method indicated.

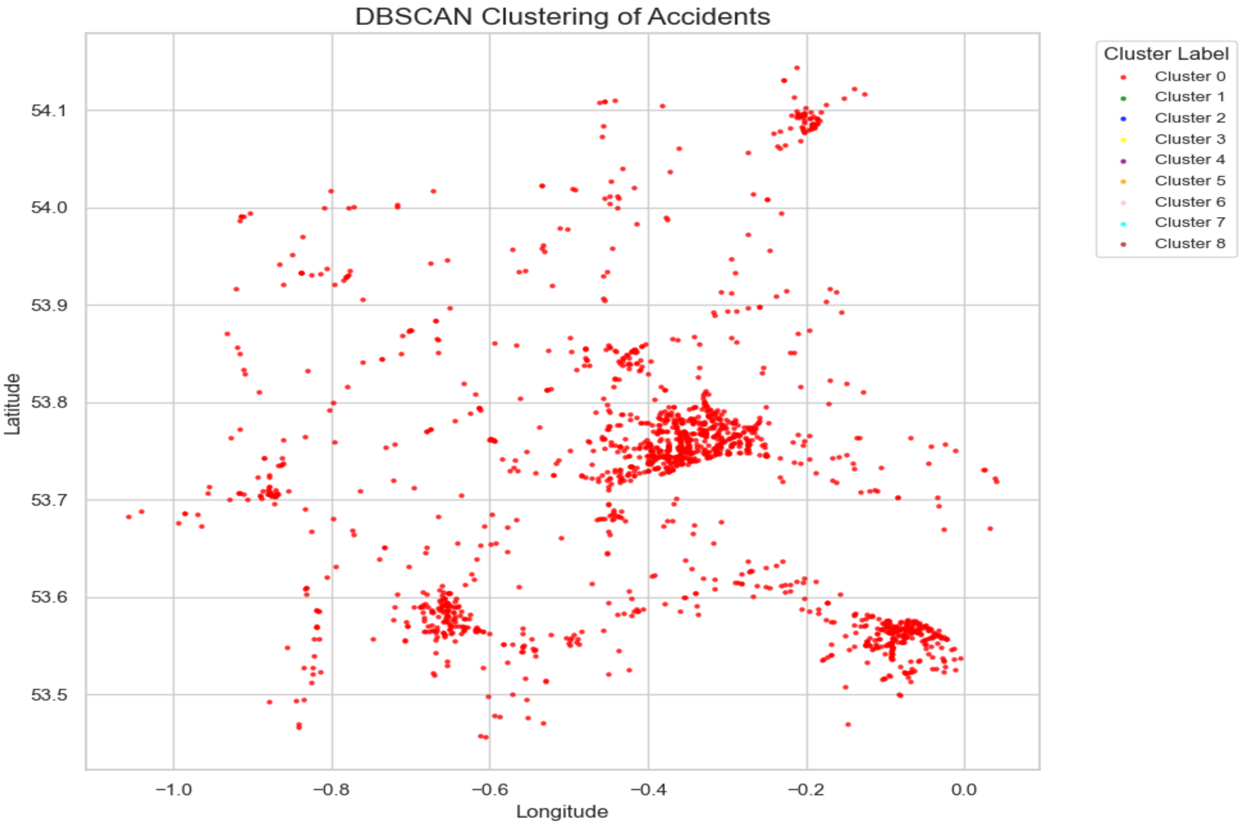


Fig 6. DBSCAN Clusters

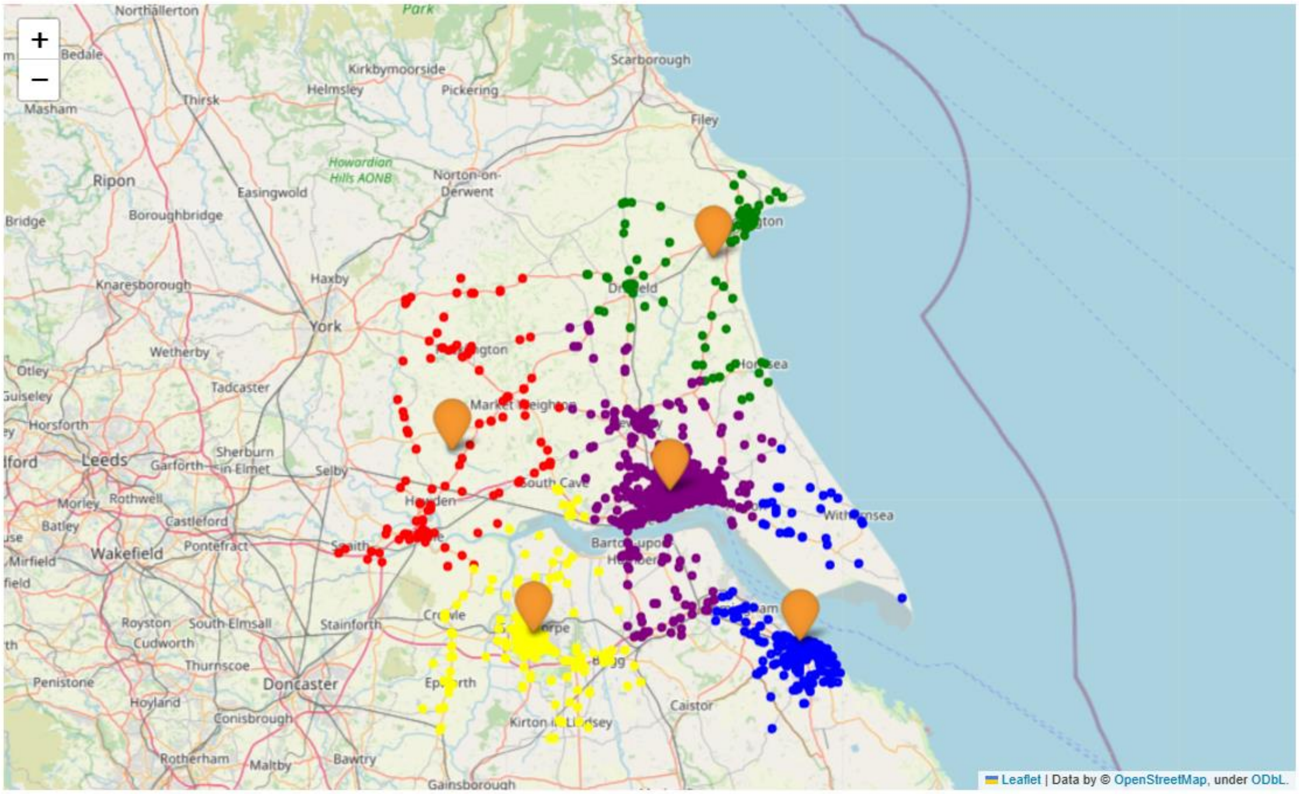


Fig 7. K-Means Clusters

To visually depict the cluster distributions on a map, the KMeans algorithm's results were chosen for representation using a folium map. The geographic coordinates (longitude and latitude) of each accident were used to position clusters on the map.

The city of Hull, Scunthorpe, and Grimsby emerged as significant clusters, reflecting high accident densities. The compactness of these clusters underscores the consistency of traffic patterns and highlights the need for targeted interventions to enhance road safety

Additionally, two other clusters, encompassing Bridlington and Goole, were identified, although they were relatively less dense. The clusters observed in Bridlington and Goole reveal unique traffic dynamics. Bridlington, being a coastal resort town, experiences fluctuating traffic patterns influenced by tourism and leisure activities. Similarly, Goole, with its industrial port operations, has specific traffic characteristics. Despite their distinct features, the presence of clusters suggests localized accident patterns.

Sparse Representation and Implications: The sparse representation in areas such as Howden, Pocklington, Driffield, and Hornsea reflects the lower accident frequency in rural or less populated regions. This affirms our understanding that areas with lower traffic volumes and simpler road networks generally experience fewer accidents. While these areas might have fewer incidents, their inclusion in the analysis is crucial for a comprehensive understanding of the region's accident landscape.

Outlier Detection

This stage was initiated by integrating the Accident, Casualty, and Vehicle data to create a comprehensive dataset that consolidated relevant information. This involved selecting the most severe casualty for each accident and subsequently merging it with the relevant vehicle data. By ensuring that duplicates were removed and the data was refined, the foundation was laid for a more accurate analysis.

Two prominent outlier detection methods were employed: Isolation Forest and Local Outlier Factor (LOF).

Both were applied to assess the potential presence of outliers within the numeric columns. With a contamination rate set at 5%, the algorithm aimed to identify data points that exhibited unusual characteristics compared to the majority. The resulting prediction identified two categories: inliers (1) and outliers (-1). The analysis revealed that out of the total entries, 4560 were classified as outliers.

Additionally, clustering analysis using DBSCAN resulted in a separate cluster represented by the label -1, signifying potential outliers. This cluster encompassed 91185 entries.

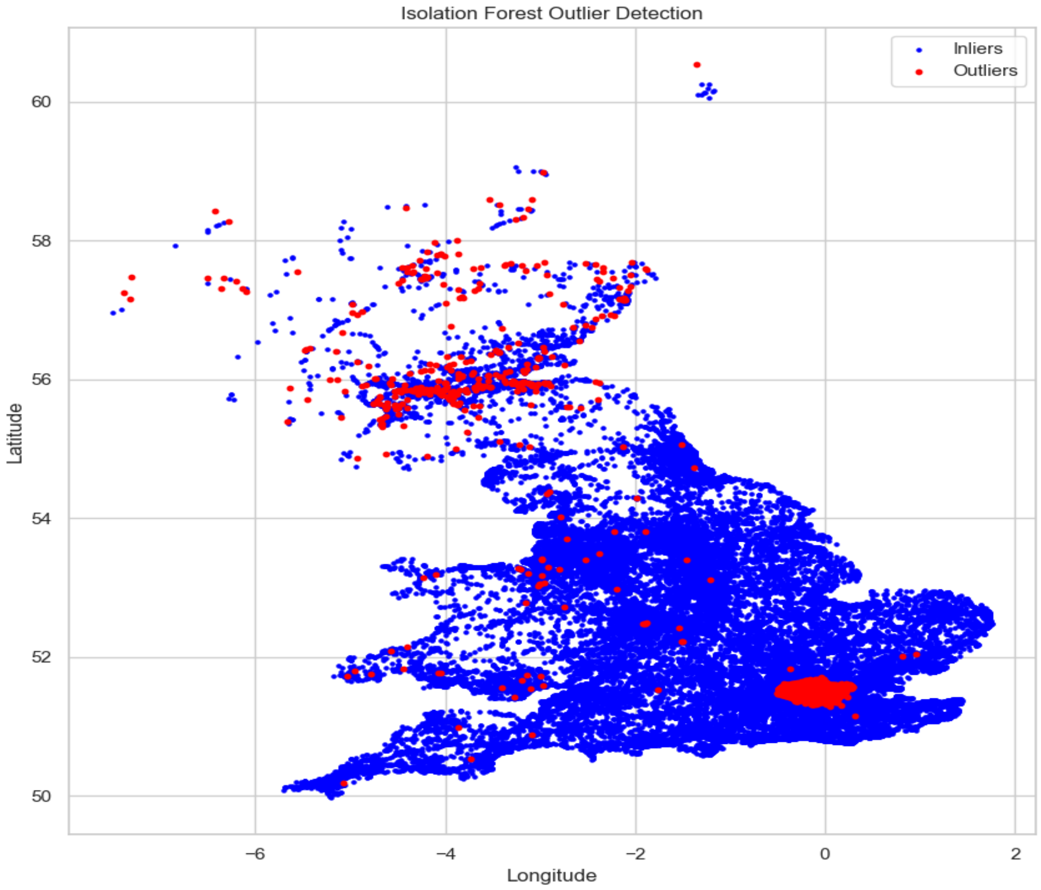


Fig 8. Outliers (Isolation Forest)

- **PREDICTING FATAL ACCIDENTS (MODEL)**

Feature Selection and Balancing:

Before building the classification model, several pre-processing steps were undertaken to optimize the dataset. Feature selection was employed to identify the most relevant variables for predicting accident severity.

To ensure balanced data for training the model, under-sampling was applied. This process involved creating a balanced dataset by reducing the number of non-fatal accidents. This balanced dataset was then utilized for further analysis.

Model Performance Evaluation:

Two classification models, Random Forest and Gradient Boosting, were developed and evaluated to predict fatal injuries in road traffic accidents.

Random Forest Classifier:

The Random Forest classifier demonstrated an accuracy of approximately 71% on the test dataset. The precision, recall, and F1-score for predicting fatal injuries were 0.68, 0.72, and 0.70, respectively. The feature importance analysis revealed that 'speed_limit' was the most influential feature, contributing significantly to the model's predictive power. Other impactful features included 'vehicle_manoeuvre', 'vehicle_leaving_carriageway', and 'vehicle_type'.

Gradient Boosting Classifier:

The Gradient Boosting classifier demonstrated an accuracy of approximately 70% on the test dataset. The precision, recall, and F1-score for predicting fatal injuries were 0.67, 0.73, and 0.70, respectively.

The developed classification models showed promising performance in predicting fatal injuries sustained in road traffic accidents. The 70 and 71% F1-score accuracy values suggest that the models are capable of providing valuable insights into identifying accidents with fatal outcomes.

However, it's important to acknowledge that the accuracy of the models could be further enhanced through continuous refinement, feature engineering, and perhaps incorporating more advanced techniques.

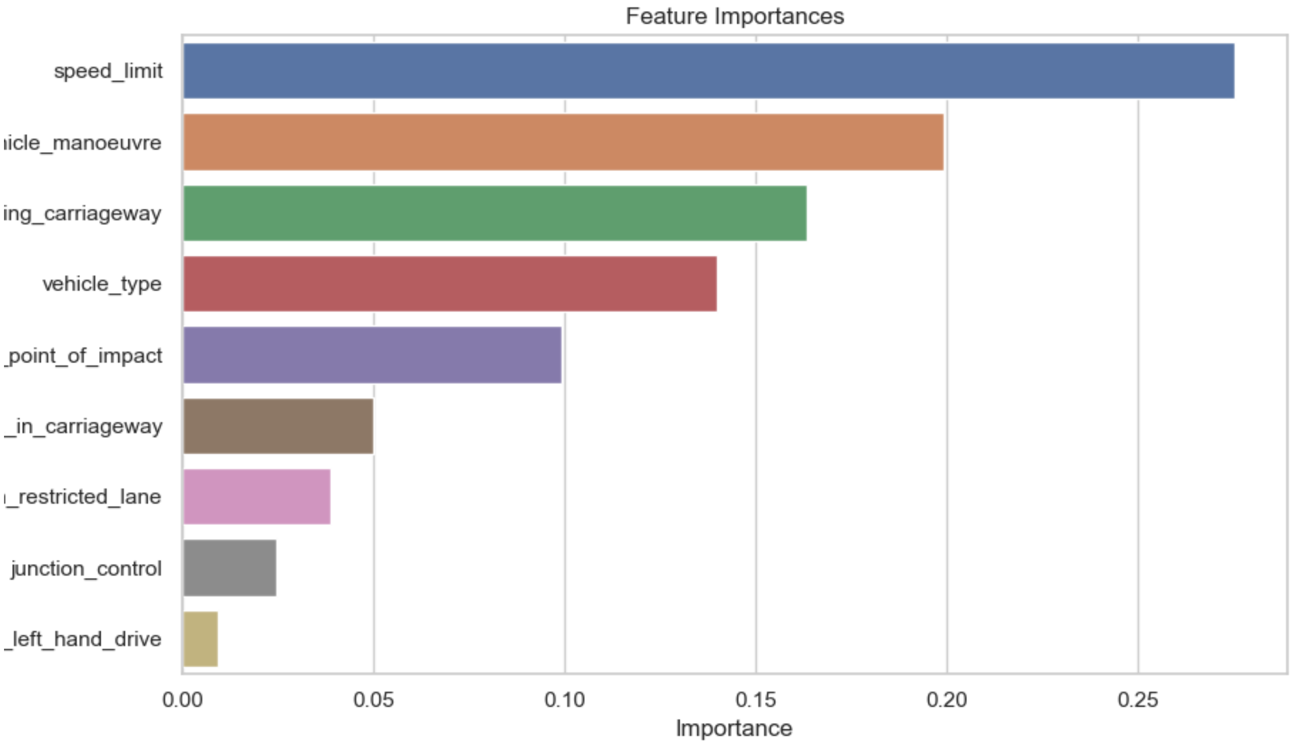


Fig 9. Model Relevant Features

- **RECOMMENDATIONS**

Implement targeted road safety measures during peak hours, particularly from 3 PM to 6 PM. Focus on improving traffic management at critical intersections and enhancing visibility during these times. Launch awareness campaigns that emphasize safe driving behaviours and pedestrian precautions to curb the higher accident frequency during rush hours.

Prioritize pedestrian safety by improving pedestrian crossings and implementing stricter enforcement around schools and busy pedestrian areas. Develop educational programs that raise awareness among both pedestrians and drivers about safe practices, especially during school start and dismissal times.

Develop custom road safety strategies targeting high-risk days, specifically Thursdays and Fridays.

For the areas in Humberside, for densely populated urban centres, allocate resources to improve infrastructure, enhance visibility, and promote responsible driving behaviours. In tourist-centric areas like Bridlington, introduce measures and enhanced signage during peak seasons to ensure the safety of both residents and visitors.

Focus enforcement efforts on curbing behaviours linked to severe accidents, such as speeding in specific zones or under specific conditions.

REFERENCES

1. Department for Transport. (2021). STATS19 road accident injury statistics – report form. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/stats19-forms-and-guidance [Accessed 13 Aug. 2023].

2. Department for Transport. (2022). Reported road casualties in Great Britain: notes, definitions, symbols and conventions. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions [Accessed 13 Aug. 2023].