ACCIDENTAL DATA ANALYSIS OF 2020

Coursework: Big Data & Data Mining

SHAON BISWAS
202251449

1. Introduction

Hundreds of road accidents occur each day, and it is important for the government agencies to keep checking the pattern of these accidents, to improve the traffic management systems, generating the rules and improving the facilities to reduce the number of accidents. In this report my goal is to analyse the accident_data_v1.0.0_2023.db database to generate some recommendations for the government agencies.

In this report I analyse the pattern of accident during 2020, with a goal to identify if accident occurred in a particular day, particular time of a day. I have also tried to find answer if there any specific time motorbikes accidents occurred, which time most accidents occurred when pedestrian envolved. I also check if there is any association rules on accident severity using the Apriori Algorithm.

Further I have checked the accident cluster of Kingston upon Hull, Humberside, and the East Riding of Yorkshire etc. area, which revealed 5 accidental clusters where accidental severity is almost identical, but difference observed in average number of vehicles, average number of casualties, average speed limits and road distribution of the road type. The outlier detection techniques are used to realise to remove outlier from the dataset. Finally, I have develop classification model in combination of decision tree and random forest classification and generate some recommendation for the government agencies.

2. Analysis

2.1 Data Cleaning

Data cleaning is an essential and fundamental step in the early stages of any machine learning project. As highlighted in the referenced paper, it holds significant importance in the overall data analysis process. This crucial step aims to eliminate inaccurate or erroneous data from the dataset, ensuring its integrity and reliability (Lee, et al., 2021).

Since in this data analysis project I am using the few machine learning algorithms, to get the best results, I like to perform some data cleaning at the initial stage. Obviously this database is SQL based relational database with four tables named accident, vehicle, casualty and LSOA table. Firstly I have connected the accidental table and coverts to panda data frame for further analysis. I found some columns have NaN values, which I have replaced with the mode of the remaining the data of the columns. In this table I have found significant amount of column with the -1 value, which is not a valid data point. In most case I have tried to clean the -1 with the mode of valid data points. In defining the valid data points I have use the stats20 pdf documents attached with this assignment specification as well as the department of transport website (Department of Transport, 2022).

Other than using the mode to replace (-1) I have also used some other methods in removing the (-1). For example, to replace the negative value in the *Local_authority_district* I have found there is common pattern of *LSOA_of_accident_location* and the *Local_authority_district* columns. In most case the same *LSOA_of_accident_loca*tion have the same *Local_authority_district*. Therefore I have replaced the (-1) in this column with the mode of common *LSOA_of_accident_location*.

In addition, to mitigate the negative one of *LSOA_of_accident_location* I have use ***KNN-based classification model*** to replace the (-1) with a valid LSOA of accident location. Since I did not find and common pattern to replace this (-1), it build the model KNN model, setting *'location_easting_osgr', 'location_northing_osgr', 'longitude', 'latitude', 'local_authority_district'* column to predict the target column of *LSOA_of_accident_location.*

Like the accident column I have cleaned few column of vehicle and casualty columns mostly by using the mode of valid data points. I got the valid data points for the state20 pdf file.

After cleaning the dataset I have created the new dataframe accident_df_2020 filtering from the accident table, with all accident occurred in 2020.

2.2 Q1. Time of Major Accident Occurrence

To get the most significant time of accident occurrence, firstly I have converted the date column to datetime objective with specific format (dd:mm:yyyy). Secondly convert the time to datetime objective as (hh:mm) format. Then combine the two datetime objects to a string value. Later I have defined Monday as 0 (1st day of the week) and Sunday as 6 (last day of the week) and generate the visualisation.

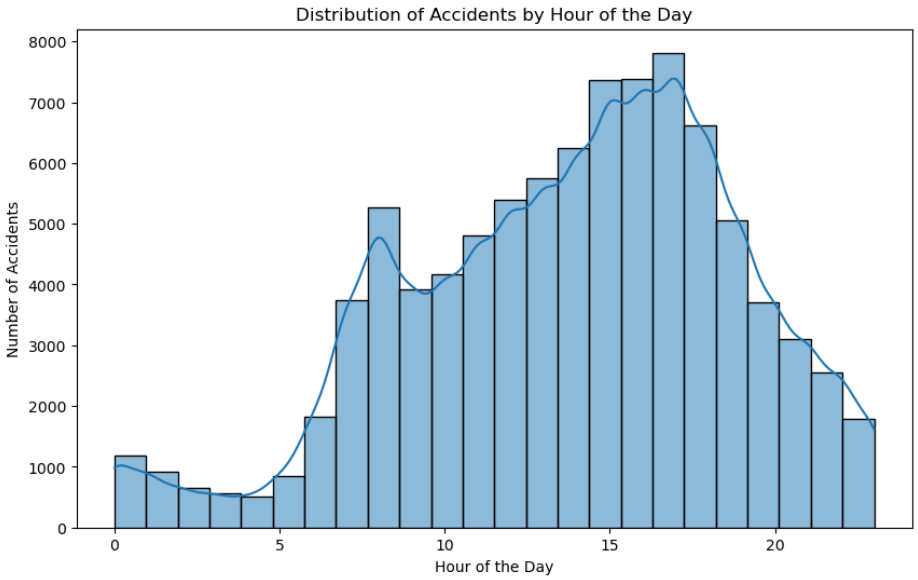


Figure 1: Hourly Accident Counts in 2020

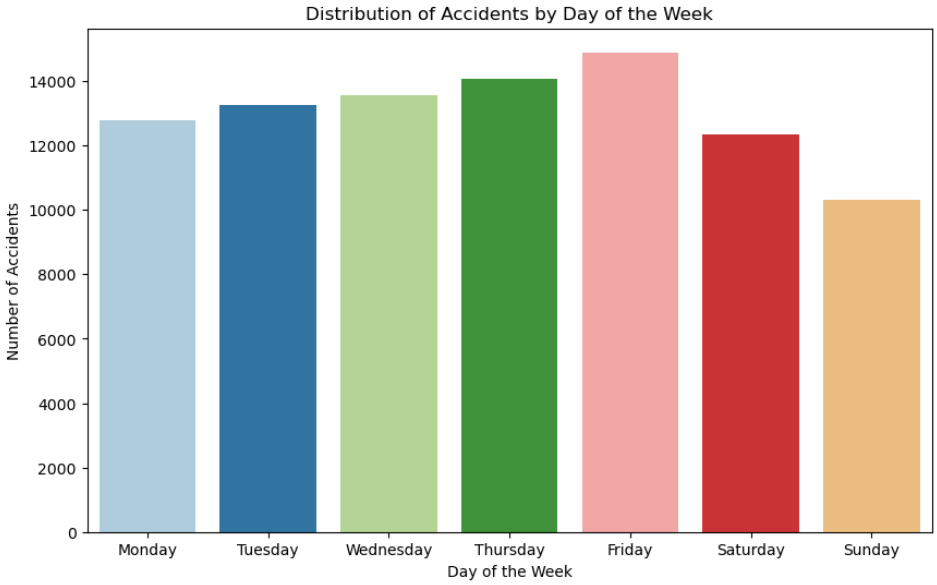


Figure 2: Accident Count by Day of Week

The figure 1 and 2 showing the most accident occurred on Friday on 17th hour. This analysis support the report of national accidental report from national accident helpline (Mehmood,

2020). 17:00th hour is time when most people heading home from office at the weekend. As people generally excited on the weekend eve, therefore most accident occurred during this hour.

2.3 Q2. Time of Accident Occurrence by Motorbikes (*Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc*)

In generating this analysis firstly I have joined the accident_df_2020 dataframe with the vehicle_df dataframe. I have create the vehicle_df by connecting vehicle table with pandas dataframe. Then using the same logic I have found that the most motorbikes accident of these three types of motorbikes occurred on Friday in same 17:00th hour. However vehicles accident incidents are significantly low in other days. This significs that the accident occurred due to the over excitement in the weekend eve.

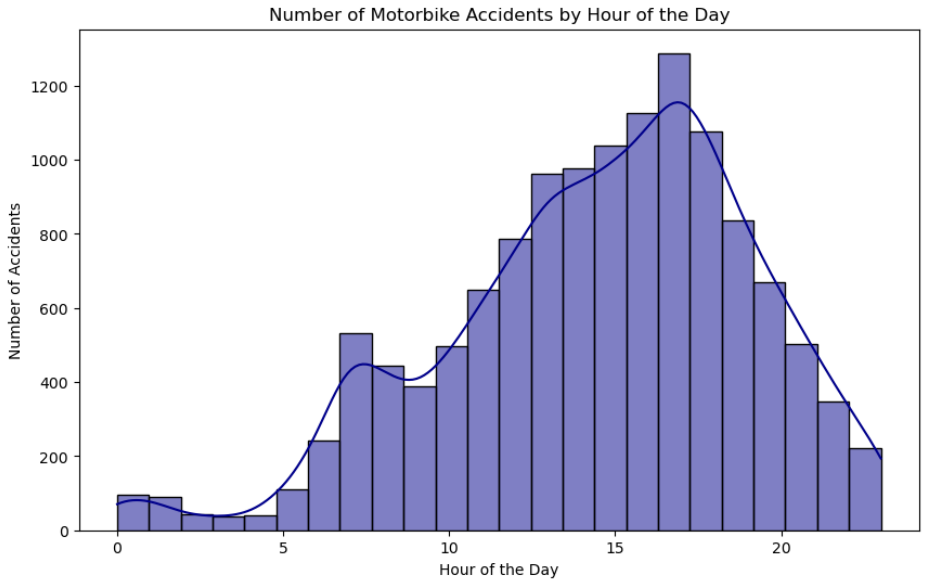


Figure 3: Motorbike Accident by Hours in 2020

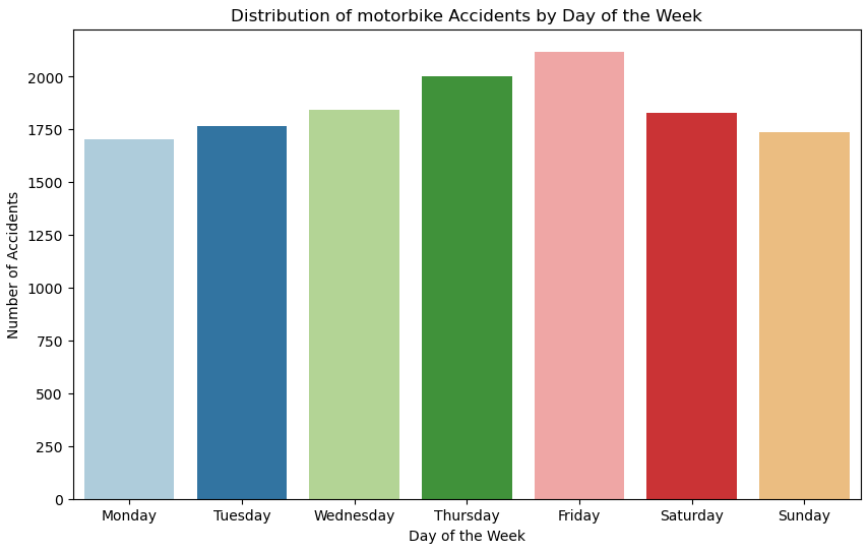


Figure 4: Motorbike Accidents by Day of Week in 2020

2.4 Q3. Time of Accident Occurrence with Pedestrian Casualty

Like the previous analysis, for this analysis I have merged the accident_df_2020 with the casualty_df (generated by connected casualty table with pandas dataframe). Then I have filtered casualty_class == 3, to filter the accident incidents with pedestrian's involvement. Later using the same datetime format have found most incidence occurred on Friday as usual, but in this case the most accident occurred on 15:00th hour. My analyse also matched with national statistics published by the department of transport on 29 September 2022 (Department of Transport, 2022).

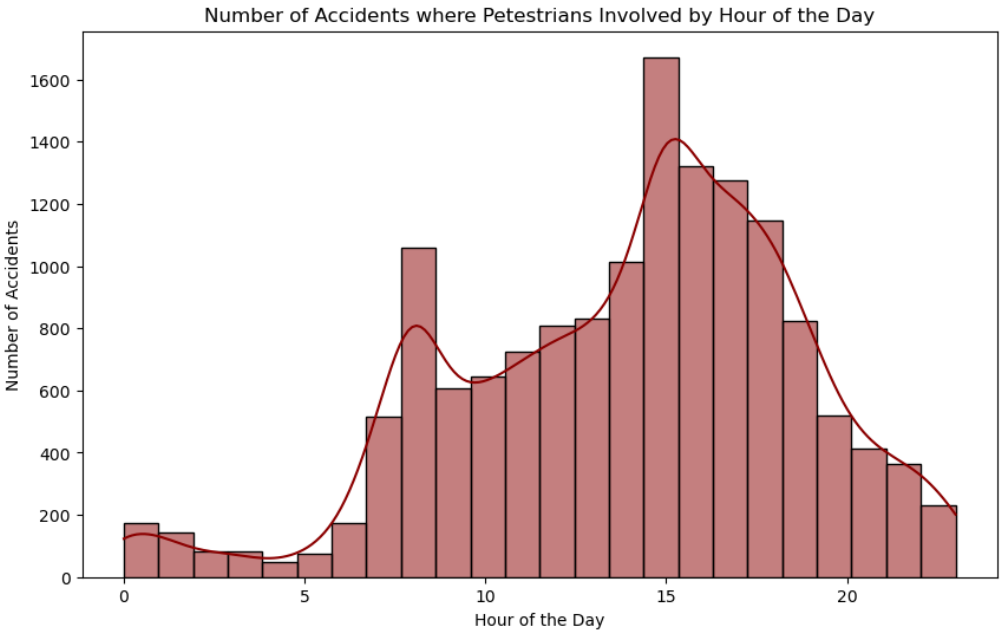


Figure 5: Accident by Hours with Pedestrian Casualty

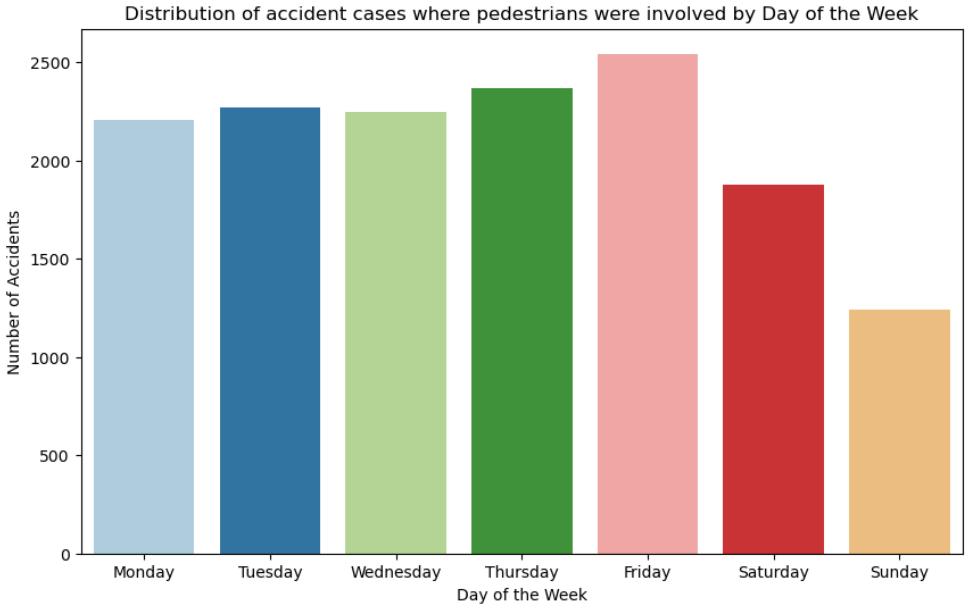


Figure 6: Accident by Day of Week with Pedestrian Casualty

2.5 Q4. Impact of Selected Variable on Accident Severity

Apriori Algorithm is an association rule mining method that is used to identify patterns associated with the and observation. Researchers such as John & Shaiba (2019), Avinash et al. (2020) discovered that driver age, accident hour, alcohol impairment, and vehicle type have a strong correlation with accident severity. Thus, I am examining whether variables such as speed limit, number of vehicles, light conditions, weather conditions, road type, vehicle type, age of driver, age of vehicle, vehicle maneuver, driver's travel purpose, and hours are associated with the severity of accidents. Due to the absence of alcohol in the dataset, I am unable to test the previous findings by John and Shaiba (2019).

For this analysis, I have used minimum support of 10%, so as to employ the most pertinent association rules. This association rule reveals significant correlations between driving conditions and accident severity. Light_1, which has been observed 129 times, and weather_1, which has been observed 115 times, are the primary factors preceding the event in question. The weather has a correlation level of 78.09%, while specific vehicle types have a correlation level of 46.98%. Certain weather conditions are associated with accident severity 2, with a level of confidence of 20.67% and a lift value near to one, indicating a weak correlation. The rules with high confidence, such as the rule (speed_30, light_1) resulting in accident_severity_3 with a confidence of 81.99%, emphasize the significance of speed and light in contributing to accident severity. Notable is the high lift of 6.5444 between specific vehicle types and maneuvers, and the perfect Zhang's metric scores of 1.0, indicating strong relationships between factors like vehicle type, light, and accident severity. The analysis underscores the complexity of factors influencing accident severity but suggests that some relationships might not be highly significant.

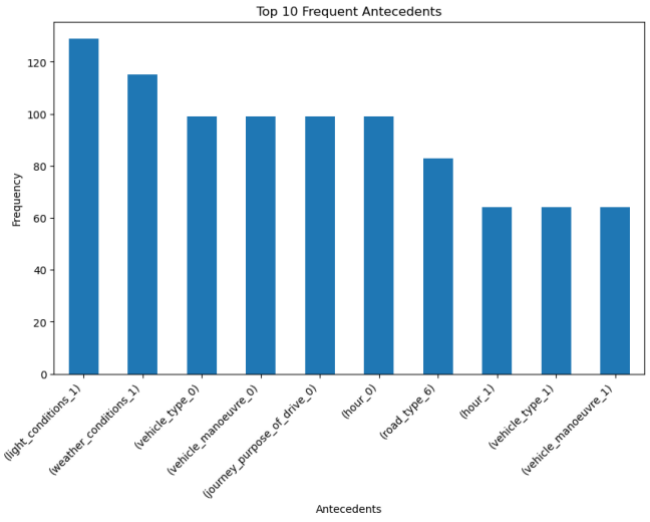


Figure 7: Top Antecedents in Association Rules

This bar plot depicts the primary antecedents within the association rules, emphasizing their frequencies within the dataset. The analysis reveals that light_1 is the most frequently occurring factor, followed by weather_1, as well as other factors associated with vehicle type, hour, and road type. The visualization presented in this study highlights the significance of specific driving conditions and their potential impact on the severity of accidents.

2.6 Q5. Accident Cluster analysis in Kingston upon Hull, Humberside, and the East Riding of Yorkshire

In generating this analysis, I have filtered out the police force to 16, which represents the selected location. K-mean has a higher Silhouette Score than other clustering techniques, which I used to choose it as the cluster analysis technique.

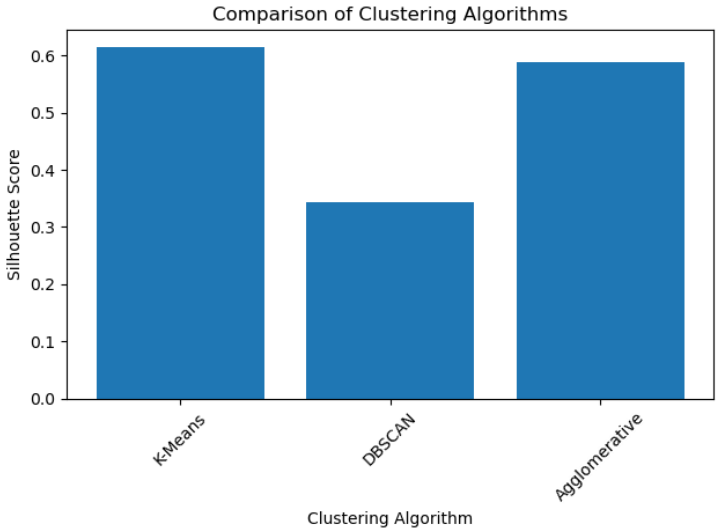


Figure 8: Comparison of Clustering Alogorithms

The Silhouette Score (SS) also determines that five clusters will maximize the SS, and I generate the five clusters based on the location of accidents. Silhouette Score is an effective way to determine the value of k in clustering algorithm (Shahapure & Nicholas, 2020).

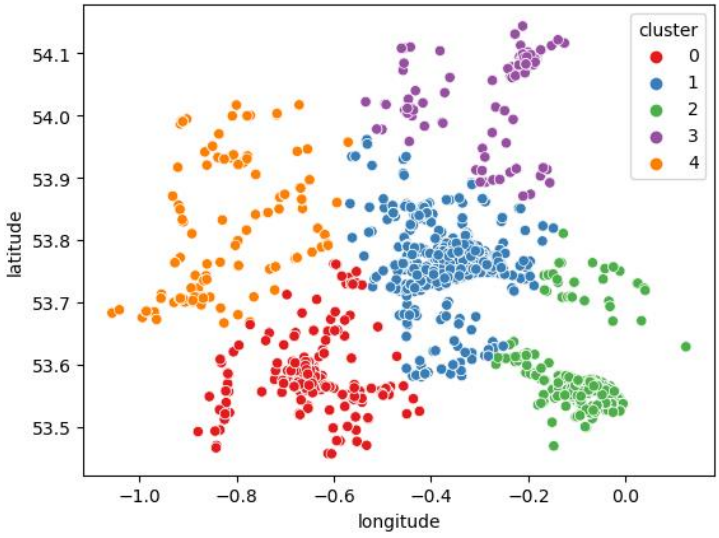


Figure 9: Location Clusters

The presence of five location clusters indicates a close similarity in accident severity (classified as "Slight") and the number of vehicles involved in each accident. Nevertheless, there is a clear distinction between the average number of speed limit violations and the average number of casualties. Furthermore, the cluster analysis has indicated that a significant proportion of incidents within this geographical area took place on roads with single carriageway, as denoted by road type 6.

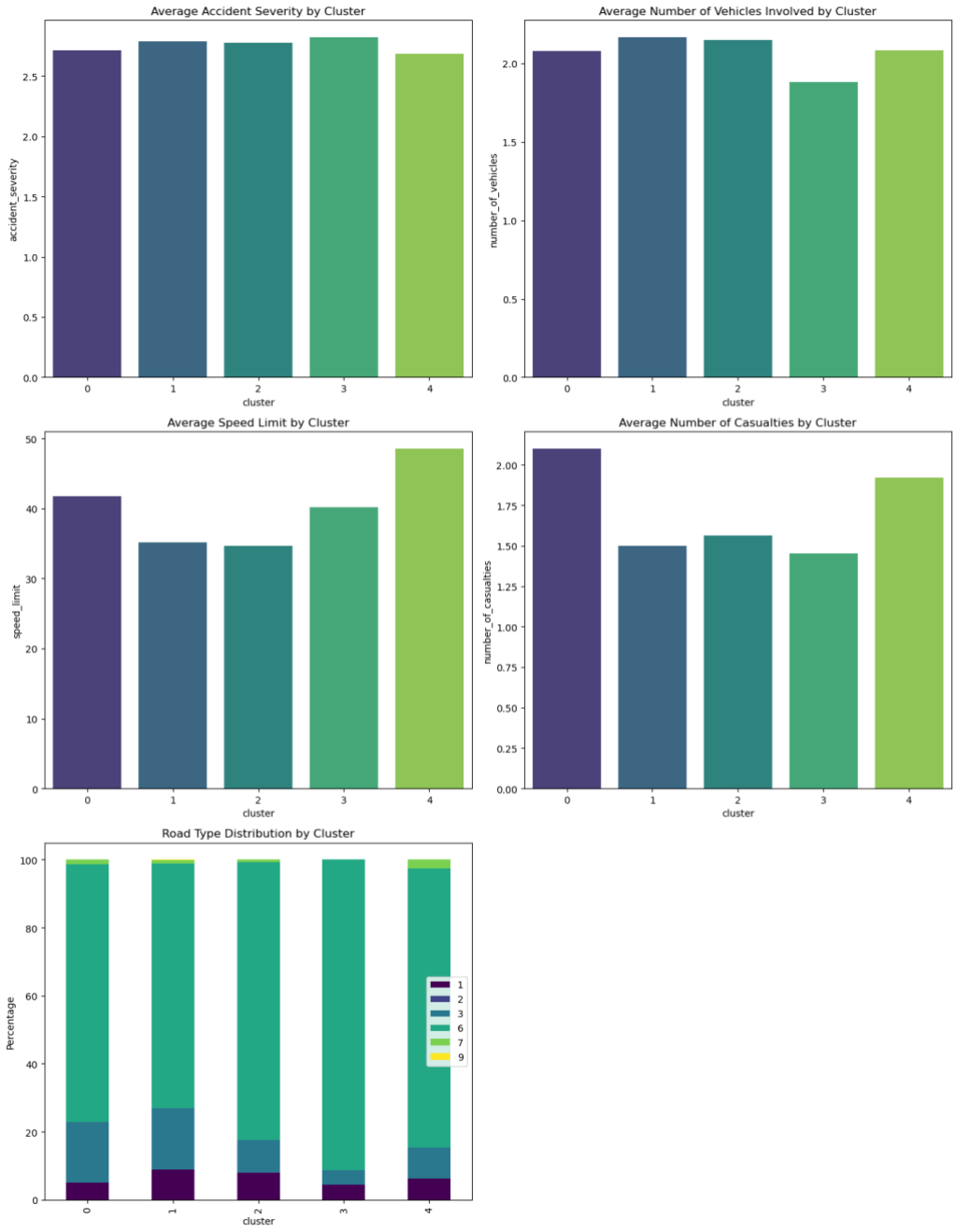


Figure 10: Accident Cluster's Behaviour Analysis

2.7 Q6. Outlier Detection

In this big dataset we have many variables. In this analysis I have tried to find out the outliers in different variables using various techniques. In age_of_driver and age_of_vehicle we have multiple negative value which I cleaned in the data cleaning state. The Z score, Interquartile Range (IQR) are the two effective way of outlier detection. These techniques showing the

both age_of_driver and age_of_vehicle have few outliers, which are data point over 90s in the age_of_driver and over 37 for age_of_vehicle.

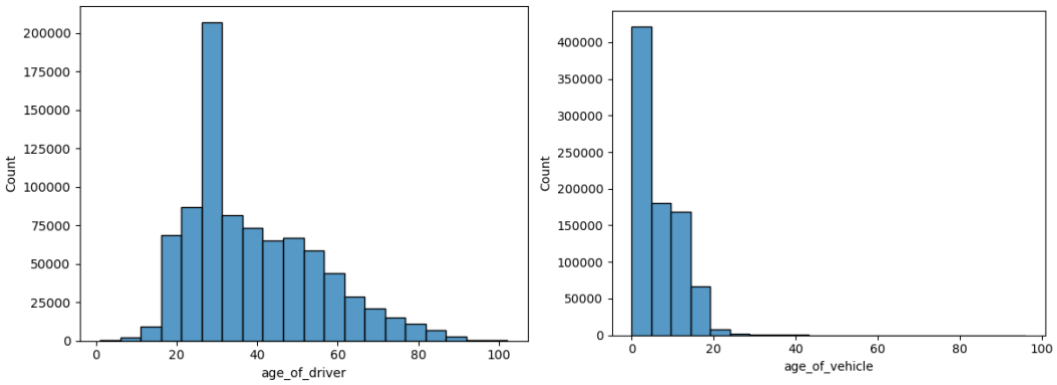


Figure 11: Age of driver and vehicle before outlier detection

Local Outlier Factors (LOF)

Local Outlier Factor (LOF) is an algorithm for identifying anomalies or outliers in a dataset. The contamination parameter indicates the percentage of outliers in a dataset. It is used to determine the distinction between inliers and outliers. A higher contamination value will produce more outliers, while a lower contamination value will produce fewer outliers.

552 data points have been identified as anomalies based on the Local Outlier Factor (LOF) algorithm and a contamination value of 0.0025. The remaining 219,883 points of data are outliers. There are 80.43 percent rural outliers and 19.56 percent urban ones.

In contrast, 64.38 percent of Inliers are found in urban areas. 35.61 percent of the population resides in rural locations.

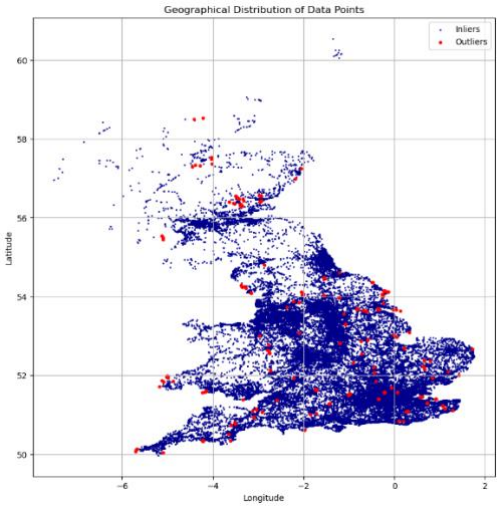


Figure 12:Local Outlier Factor

Isolation Forest

The Isolation Forest is a machine learning algorithm used for outlier detection. It detect outlier with set of contamination values set at 0.05 point interval, and results various outlier at various contamination. In this data set I used the isolation forest in spatial data analysis.

Contamination Value	Number of Outliers
0.01	907
0.05	4539
0.1	9118
0.2	18236

Table 1: Isolation Forest Contamination and Outlier Counts

The visualization of the outliers based on longitude and latitude shows that they are scattered throughout the region, though some clustering can be observed. This may suggest that the outliers correspond to particular subregions or spatial patterns.

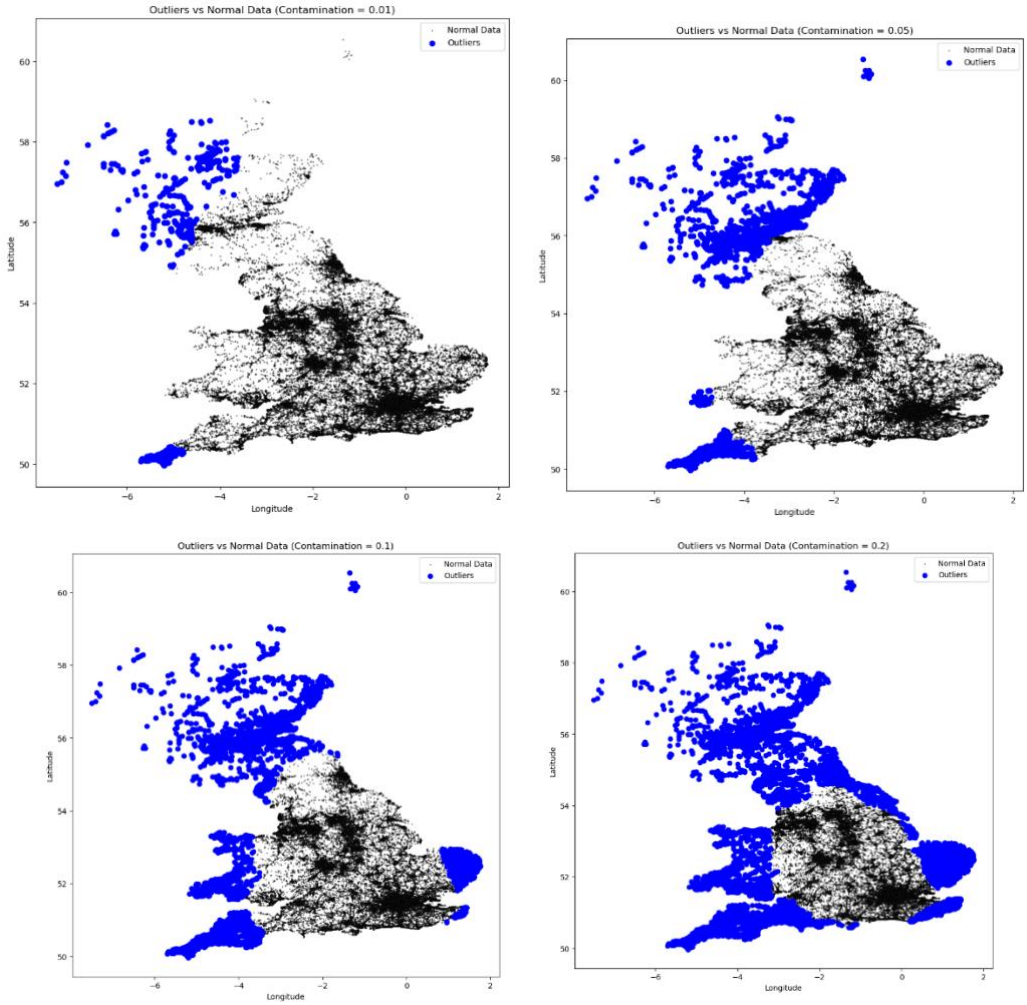


Figure 13: Contamination vs Outliner in Isolation Forest

3. Predictions

To build a classification model that can predict the fatal injuries in the accident, initially I have initially 20 selected features like speed limit, number of vehicles, weather conditions, light condition and so on. With this features I set the target as accident severity (1) where 1 represent the fatal accident.

The data distribution of accident severity shows that fatal accident represents only 1.92% of the data set. Therefore random under sampling or Synthetic Minority Over-sampling Technique (SMOTE) are used to deal with data imbalance firstly.

	Count	Percentage
3	171376	77.744460
2	44828	20.336154
1	4231	1.919387

Initially I used K-Best Algorithm to generate the feature importance.

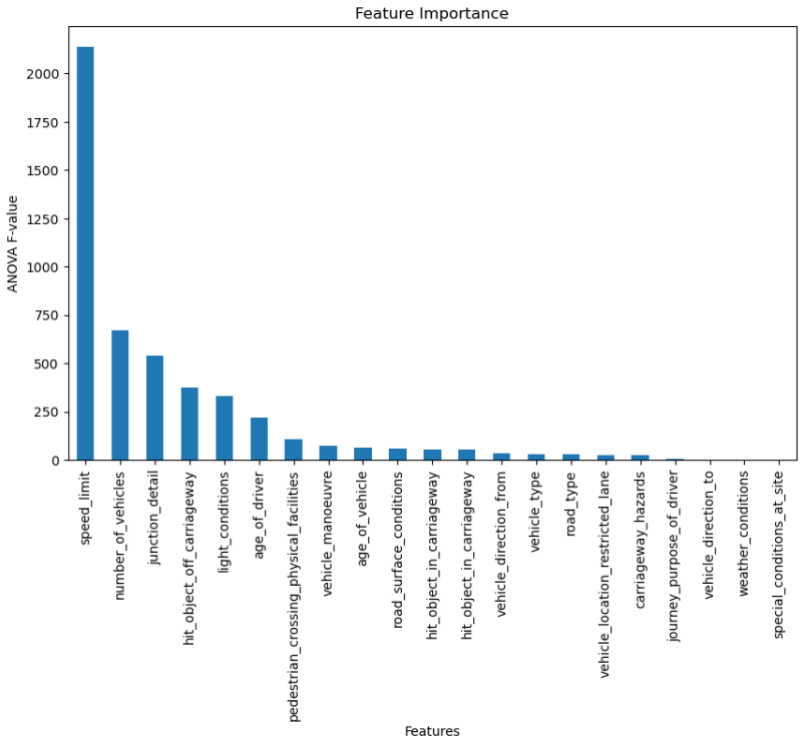


Figure 14: Feature Importance Using K-Best

Then I used decision tree classifier to predict the fatal accident. The Random under sampling and SMOTE product the decision tree accuracy 72.06 and 98.15% respectively. Hence I use the SMOTE in comparing the result with other classification techniques. The result shown that the random forest can predict with more accuracy.

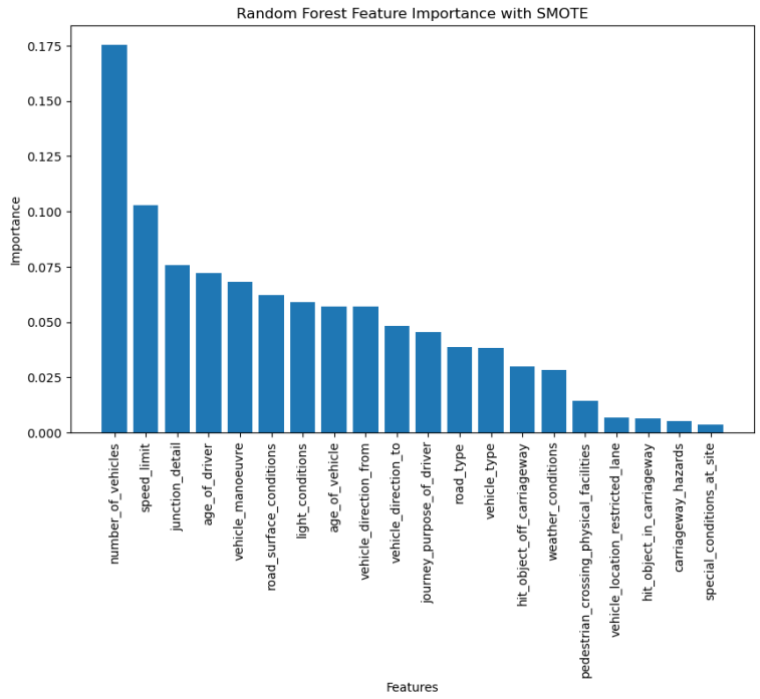
```
Decision Tree Accuracy with SMOTE: 98.15365073604464
Random Forest Accuracy with SMOTE: 99.19477396965092
K-Nearest Neighbors Accuracy with SMOTE: 93.7850159911085
Logistic Regression Accuracy with SMOTE: 71.22507768730011
```

The RF algorithm provide accuracy of 99%, but less recall score for fatal accident.

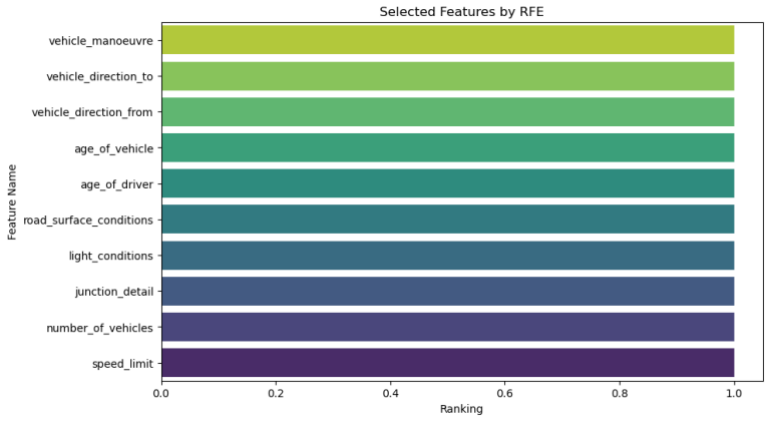
Random Forest Classification Report with SMOTE:

	precision	recall	f1-score	support
Not Fatal	0.99	1.00	1.00	43241
Fatal	0.95	0.61	0.74	846
accuracy			0.99	44087
macro avg	0.97	0.81	0.87	44087
weighted avg	0.99	0.99	0.99	44087

Thus I run the hyperparameter tuning to improve scores. The RF algorithm detect the following feature importance.



Later to detect the best feature I have used Recursive Feature Elimination (RFE) techniques to best features, which shows features like speed limit, number of vehicles, junction details, light conditions, road surface conditions, age of vehicle, age of drivers, vehicle direction from, vehicle direction to and vehicle manoeuver as the most significant parameter for fatal accident.



The confusion matrix and cross validation accuracy from standard deviation significs that the model is good to predict the fatal accident.

4. Recommendation

To enhance road safety and mitigate the occurrence of accidents, a multifaceted approach can be adopted. This involves strengthening speed limit regulations in high-risk areas, bolstering the presence of speed cameras and patrols, and initiating public awareness campaigns on the perils of speeding. Monitoring traffic patterns and congestion in accident-prone zones is crucial, with the application of traffic management strategies during peak hours. Evaluating the design and security of intersections and junctions is essential, including enhancing signage, illumination, and road markings, potentially incorporating traffic signals or roundabouts. Addressing weather-related concerns entails improving nocturnal visibility through better street lighting and promoting the use of appropriate headlights and reflective materials for vehicles and pedestrians. Consistent road surface upkeep and drainage maintenance are imperative to prevent hazardous slick conditions caused by precipitation accumulation. Considering the age factor, additional assessments and educational initiatives could be introduced for novice and elderly drivers, complemented by defensive driving courses. Encouraging regular vehicle maintenance and incentivizing the transition to safer, newer models can significantly contribute to safety. Promoting safe driving practices and optimizing road layouts for seamless traffic flow are vital, necessitating collaboration among governmental bodies, law enforcement, transportation departments, and local communities for effective implementation.

5. Bibliography

- Avinash, V., Gupta, A. B. & Mary, S. P., 2020. Analysis of Road Fatal Accidents Using Apriori Algorithm. In: N. Priyadarshi, et al. eds. *Power Systems and Energy Management.* Singapore: Advances in Power Systems and Energy Management: Select Proceedings of ETAEERE 2020, pp. 627 - 633.
- Department of Transport, 2022. *GOV.UK.* [Online]
  Available at: https://www.gov.uk/government/statistical-data-sets/reported-road-accidents-vehicles-and-casualties-tables-for-great-britain#all-collision-casualty-and-vehicle-tables-excel-format
  [Accessed 13 August 2023].
- Department of Transport, 2022. *GOV.UK.* [Online]
  Available at: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-pedestrian-factsheet-2021/reported-road-casualties-great-britain-pedestrian-factsheet-2021
  [Accessed 13 August 2023].
- John, M. & Shaiba, H., 2019. *Apriori-Based Algorithm for Dubai Road Accident Analysis.* Dubai, Procedia Computer Science, pp. 218-227.

- Lee, G. Y., Alzamil, L., Doskenov, B. & Termehchy, A., 2021. *A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance,* s.l.: arXiv: Databases.
- Mehmood, W., 2020. *National Accident Helpline.* [Online]
  Available at: August
  [Accessed 13 August 2023].
- Shahapure, K. R. & Nicholas, C., 2020. *Cluster Quality Analysis Using Silhouette Score.* Sydney, IEEE, pp. 747-748.