Improving Road Safety: Utilizing Data Analysis to Predict Accident Probabilities and Prevent Injuries

This project aims to use data analysis to improve road safety by predicting accident probabilities and associated injuries. By understanding the patterns and circumstances of accidents, we can suggest ways to minimize risks and promote responsible behavior. Road traffic accident reports provide valuable data for this purpose.

Methodology

In order to increase the accuracy of our predictions, we first conducted data preprocessing, including exploratory data analysis and cleaning. Next, we utilized the Apriori algorithm for association mining and clustering to gain a better understanding of accident causes. The data was then normalized to facilitate model construction, followed by feature selection. Finally, the Random Forest, Decision Tree, Logistic Regression, and Gradient Boost algorithms were employed to predict the severity of traffic accidents.

Data description and cleaning

In order to accurately predict accident severity, we carefully selected a dataset of 31 relevant features from the accident, vehicle, and casualty tables. Our selection process was guided by specific criteria designed to prioritize attributes with a significant impact on the prediction. The resulting dataset consisted of 220,435 entries and 31 columns. During preprocessing, we removed the 'time' feature and eliminated 22,258 duplicate entries. We also identified 29 instances of missing values in the 'longitude' and 'latitude' features, which we will replace with averages specific to police force code 63. Additionally, we will address any negative values, particularly in the longitude column, to ensure accurate data analysis. Finally, our imputation strategies include mode replacement for most features and median imputation for continuous numerical data such as "age_of_vehicle" and "age_of_driver."

Geographical Analysis of Traffic Accidents

Utilizing Folium, a geographical data visualization tool, we were able to map out the high-density accident hotspots across the UK. The results showed a significant concentration of accidents in cities such as Liverpool, Leeds, and Southampton. Interestingly, these cities also happen to be the most populous areas in the UK (Feng et al., 2020). This analysis also revealed that the majority of accidents (more than 75%) occur on single carriage roads within urban areas, as seen by the clustering of red dots in inner city areas (Fig. 1). This could be due to higher local traffic congestion compared to highways. Additionally, it was found that approximately 78.35% of recorded accidents are classified as minor, while fatal injuries account for less than 1.58% of the total accidents (Fig. 2). This data suggests that the overall impact of accidents may not be as severe as perceived. Furthermore, the data also showed that there were more male casualties and male drivers compared to females, with the age group between 26 and 35 years being the most susceptible to accidents (Fig. 3).

Fig. 1: UK accident hotspot

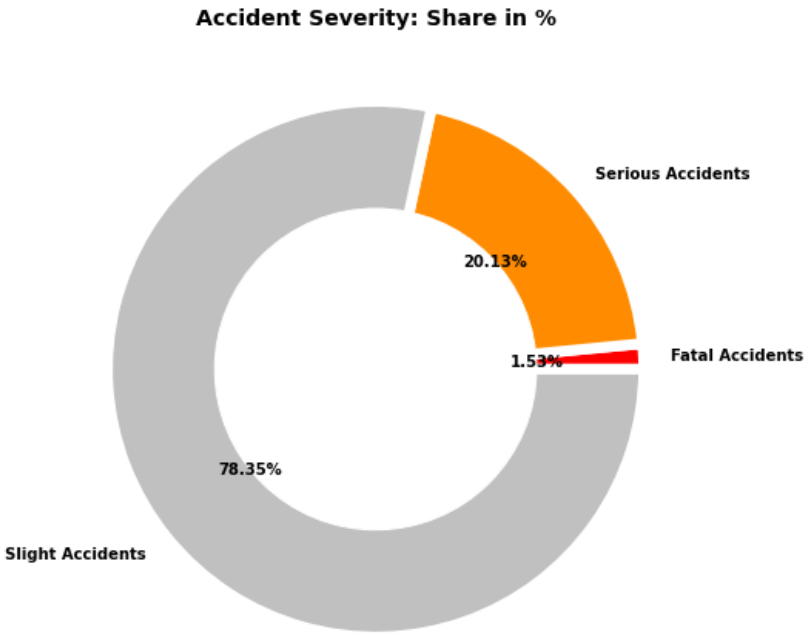

Fig. 2: UK accident severity distribution

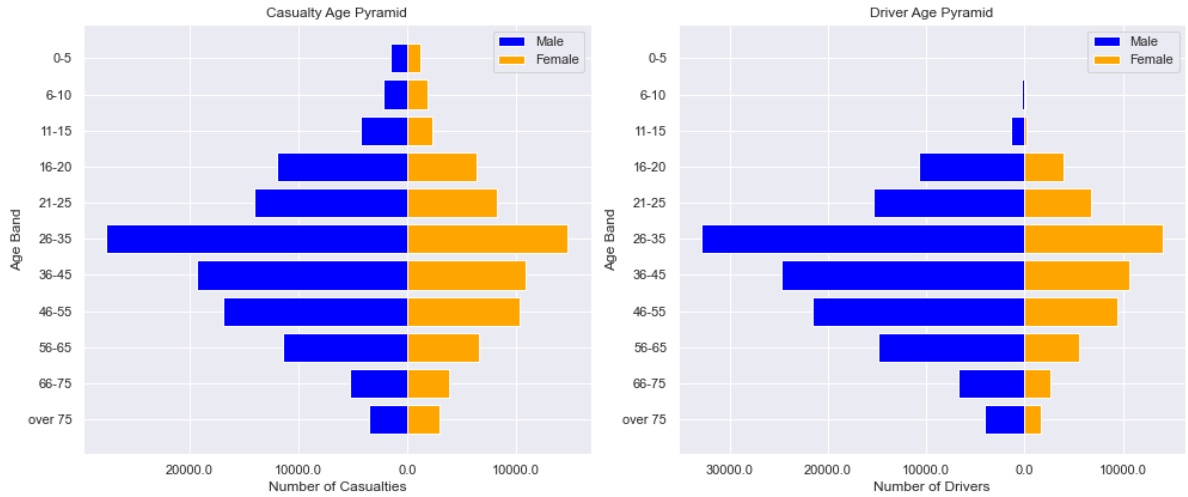

Fig. 3: Casualty and driver age pyramid

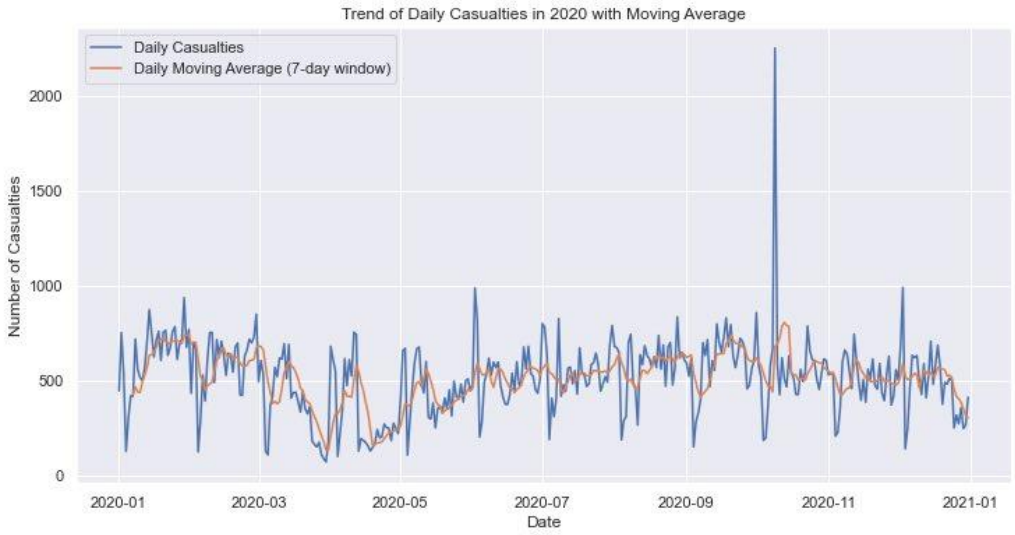

Fig. 4: Accident casualty variation during the year.

Insights from accident data analysis

Peak hours and days for accidents

The chart displays the frequency of accidents reported at various times throughout the day. The horizontal axis represents the hour of the day, while the vertical axis represents the number of accidents. The largest number of accidents happened between 16:00 and 17:00, accounting for 8.6% of the total accidents. In contrast, the time period with the fewest accidents was between 4:00 and 5:00, with only 0.6% of accidents occurring during this time (Fig. 5).

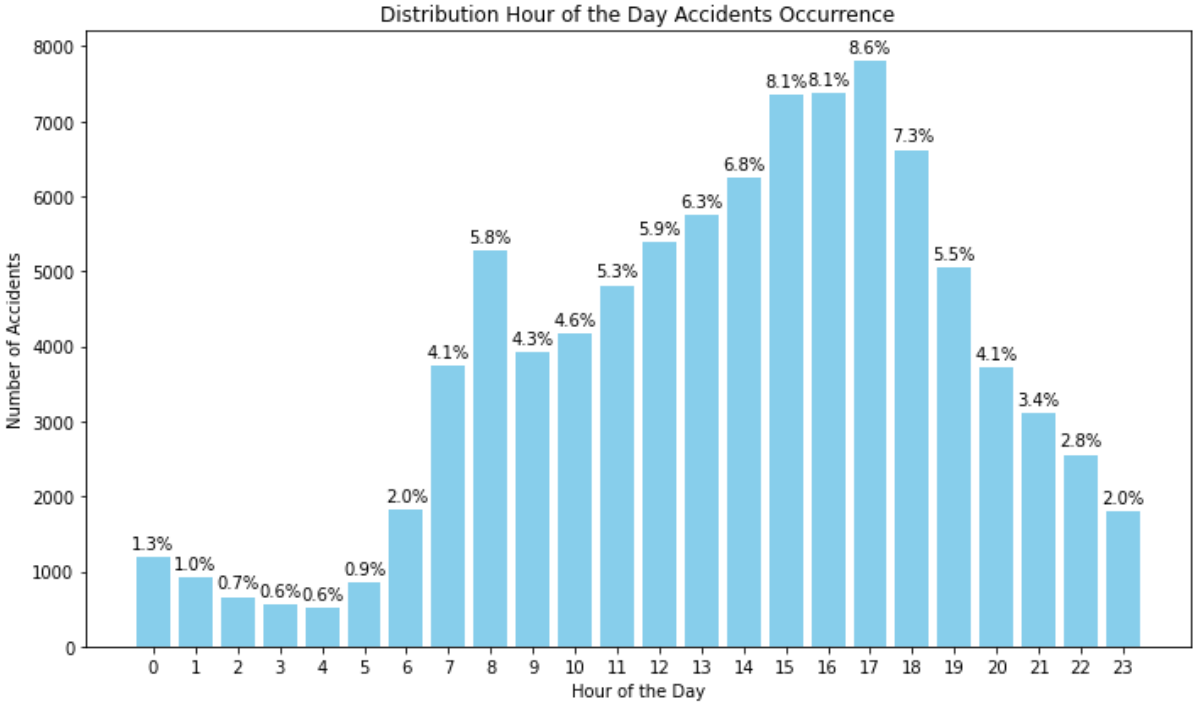

Fig. 5: Accident occurrence during the hours of a day

The bar graph depicts the distribution of accidents throughout the week. Interestingly, Friday stands out as the day with the most frequent accidents, making up 16.3% of the total. This trend could be attributed to the increased activities and excitement for the upcoming weekend. Close behind is Thursday, contributing 15.4% of the total accidents, indicating a similar pre-weekend dynamic. Mid-weekdays, Wednesday and Tuesday, also have considerable accident rates at 14.9% and 14.5%, respectively, possibly due to accumulated stress and fatigue. Monday, which reflects the start of the workweek, accounts for 14.0% of accidents. The number of accidents decreases on weekends, with Saturday at 13.5%, likely due to leisure activities, and Sunday recording the lowest percentage at 11.3%, reflecting reduced work-related travel. Understanding these patterns can help target safety measures and interventions in traffic management, especially on weekdays with higher accident risks (Fig. 6).

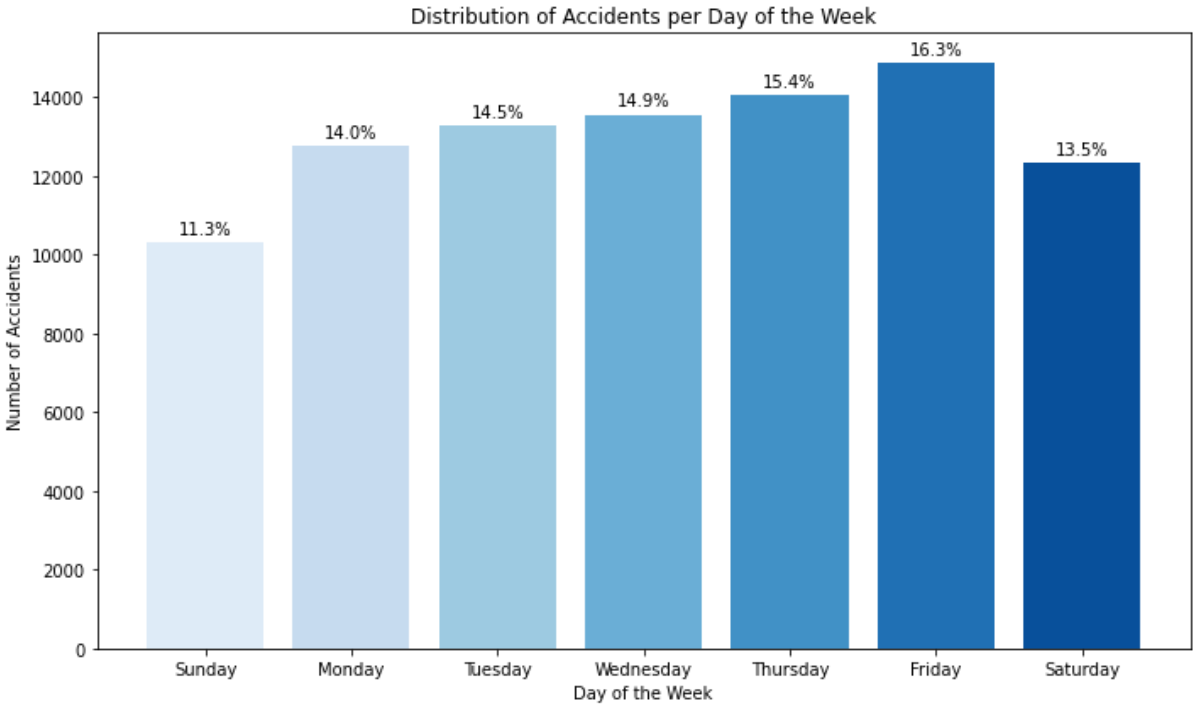

Fig. 6: Accident occurrence per weekday

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

The data on motorcycle accidents follows a similar pattern to overall accident distribution. The number of accidents steadily increases from 5:00 AM, reaching a peak at 5:00 PM with 9.9% of accidents occurring during this time. After the peak, the number of accidents decreases until it reaches its lowest point of 0.3% around 4:00 AM, before increasing again. This trend can be explained by a number of factors. The morning increase in accidents coincides with rush hour traffic, when there is a higher volume of vehicles on the road and a greater likelihood of accidents. The evening peak at 5:00 PM can also be attributed to rush hour traffic, as people are heading home from work. The decrease in accidents from the evening peak until the early morning hours is likely due to reduced traffic during nighttime, when there are fewer motorcycles on the road (Fig. 7).

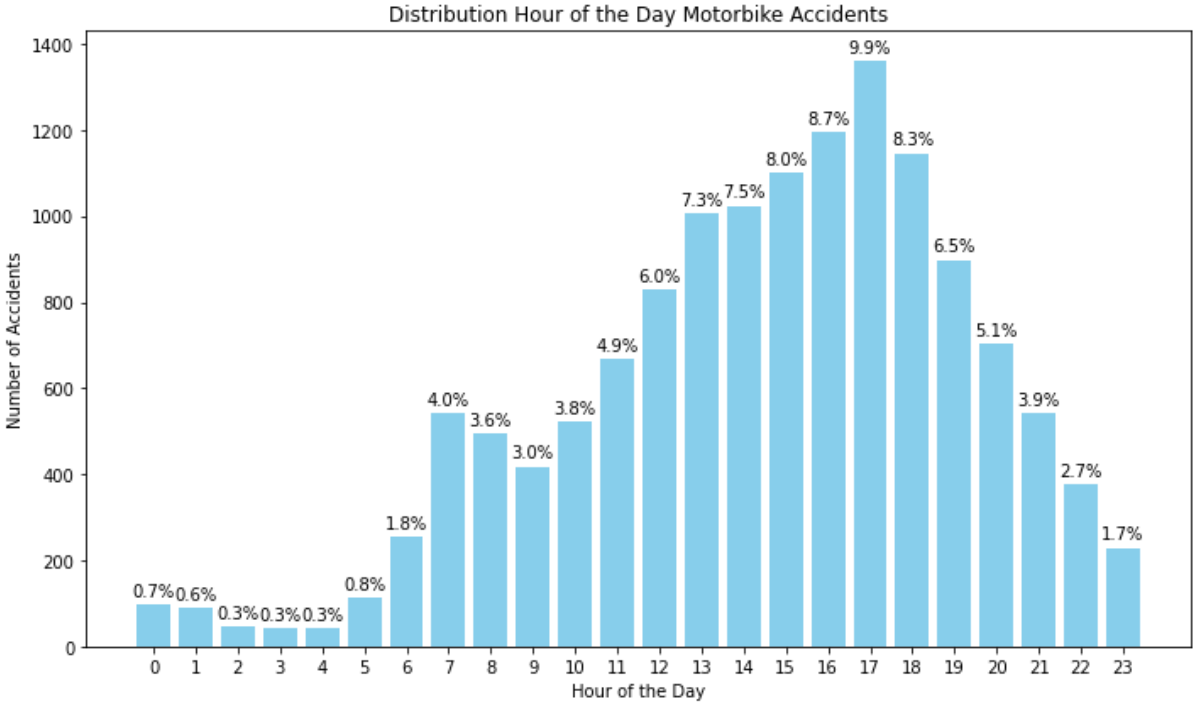

Fig.7: For motorbikes, the significant hour for accident occurrence

The barplot for motorcycle accidents by day of the week follows a similar trend to the overall accident distribution. Friday records the highest number of accidents, with 16.4% occurrences, followed closely by Thursday with 15.5% accidents. Wednesday and Tuesday show slightly fewer accidents but still exhibit increased occurrences during the middle of the week. Monday continues the trend of higher accidents at the beginning of the week, while Saturday has a lower accident count compared to weekdays. Sunday records the lowest number of accidents, with 12.8% occurrences. In summary, the analysis reveals a clear pattern of higher accident rates during weekdays, particularly on Fridays and Thursdays. Weekends, especially Sundays, see fewer accidents, likely due to reduced work-related travel and more leisure-oriented driving. These findings can inform road safety measures and traffic management strategies, emphasizing accident-prone days to enhance overall safety (Fig. 8).

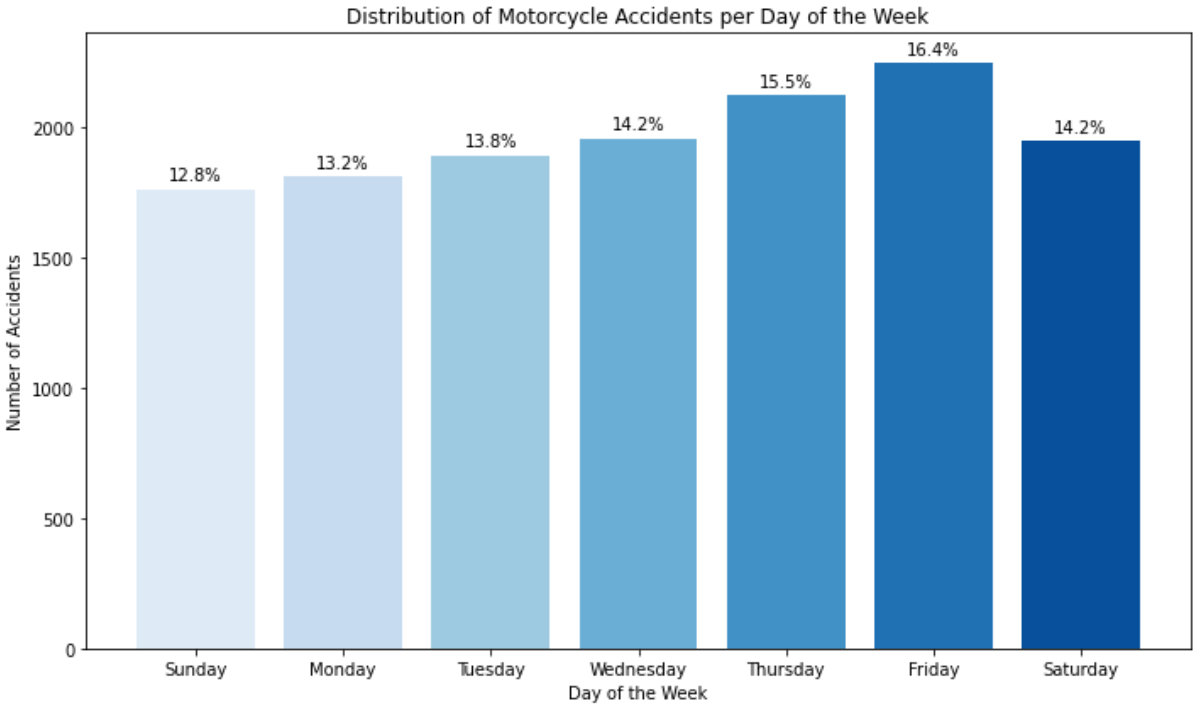

Fig.8 Accident occurrence per weekday for motorbikes

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

The data clearly displays a consistent pattern in pedestrian accidents throughout the day. The number of accidents steadily increases from 5:00 AM to 3:00 PM (15:00 PM), reaching its peak at 3:00 PM (15:00 PM) before gradually declining until 4:00 AM. This peak at 3:00 PM (15:00 PM) suggests that this time period experiences the highest number of pedestrian accidents, likely due to heightened pedestrian activity during daytime hours. The relatively lower number of accidents between 1:00 AM and 4:00 AM may be attributed to decreased pedestrian movement during these early morning hours (Fig. 9).

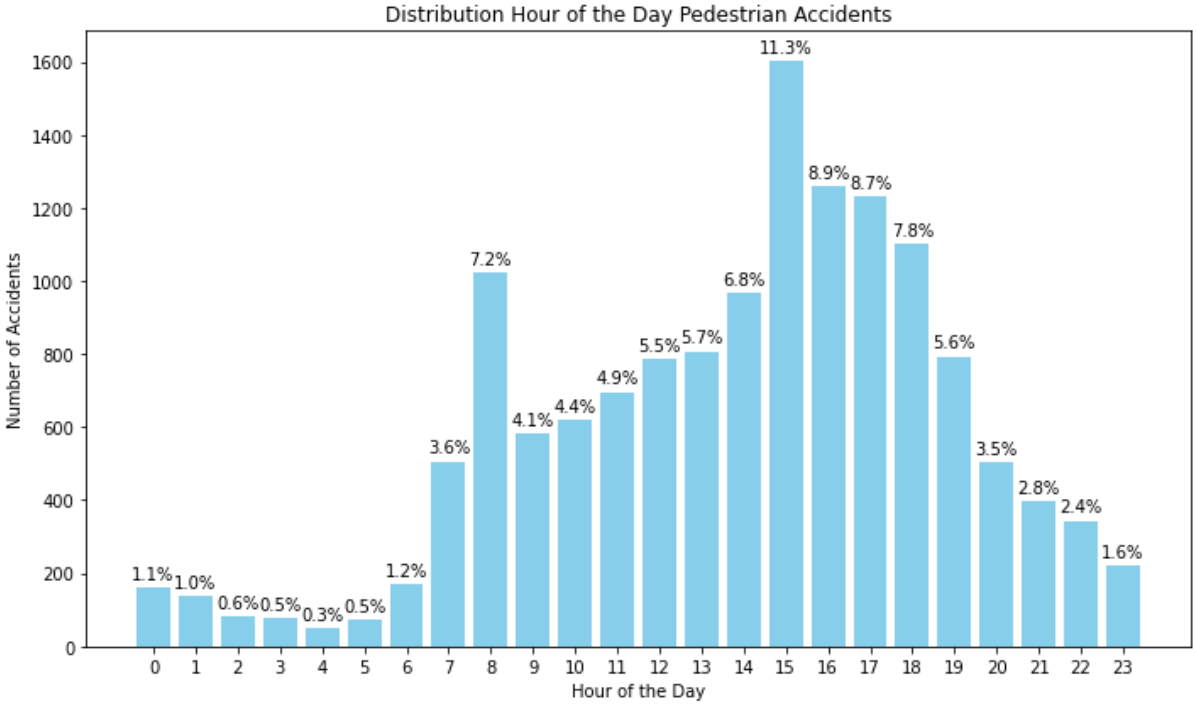

Fig. 9: Significant hours of the day for accidents involving pedestrians.

The bar plot depicting pedestrian accidents by day of the week reveals a consistent pattern similar to the overall accident distribution. According to our data, Friday has the highest number of pedestrian accidents, accounting for 17.2% of incidents. This is likely due to increased pedestrian activity at the start of the weekend. Thursday closely follows with 16% of accidents, indicating a trend of elevated pedestrian accidents towards the end of the workweek. On Wednesday and Tuesday, there are slightly lower accident counts (15.3% and 15.4% respectively), suggesting continued pedestrian involvement during the middle of the week. Monday's pedestrian accidents (15%) align with the overall trend of higher accidents at the beginning of the week, possibly due to increased commuter traffic. In contrast, Saturday records fewer pedestrian accidents (12.7%), reflecting reduced weekday work-related activity. Lastly, Sunday has the lowest count of pedestrian accidents (8.5%), likely due to decreased overall activity and fewer pedestrians on weekends (Fig. 10).

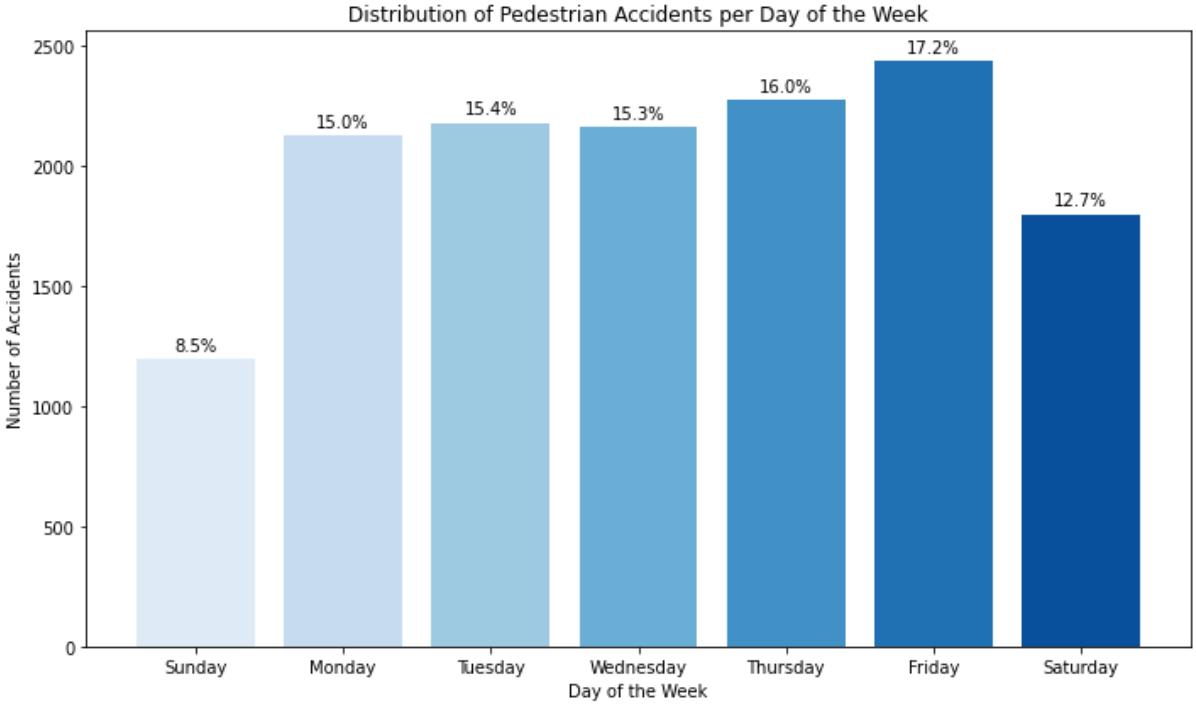

Fig. 10: Accident occurrence per weekday for pedestrian

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

The Apriori algorithm was used to analyze the impact of various factors, including weather conditions, speed limit, light conditions, urban or rural area, and road surface conditions, on accident severity. These key attributes were extracted from accident reports and one-hot encoded before being fed into the algorithm to establish association rules. This method will allow us to evaluate the relationship between these variables and accident severity.

Table 1 presents several rules indicating conditions under which slight accidents frequently occur in urban areas. Rule 1 shows that accidents in urban areas during daylight are correlated with slight accidents, with 38.9% support, 82.2% confidence, and a lift of 1.049. This suggests that most accidents occurring in urban areas during daylight result in minor injuries.

Rule 2 indicates that in urban areas with dry roads and daylight conditions, there is a correlation with slight accidents, evidenced by 30.2% support, 81.8% confidence, and a lift of 1.044. This

highlights that these conditions also frequently result in minor injuries. Rule 3 points out that urban areas with daylight conditions and a speed limit of 30 km/h are associated with slight accidents, with 28.1% support and 81.7% confidence. This implies that lower-speed urban areas tend to influence the occurrence of minor accidents. Rule 4 suggests that urban areas with daylight, dry roads, and fine weather without high winds are linked to slight accidents, showing 28.2% support and 81.5% confidence. This indicates that these multiple favorable conditions contribute to the occurrence of minor accidents. Finally, Rule 5 implies that urban areas with daylight and fine weather without high winds are correlated with slight accidents, with 31.7% support and 81.5% confidence. This underscores that these favorable conditions frequently result in minor accidents. Overall, the data indicate that urban environments, particularly under good weather and road conditions, are conducive to minor accidents.

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

Rule	Association	Support	Confidence	Lift
1	{Urban, Daylight} => {Slight}	0.389	0.822	1.049
2	{Urban, Dry_road, Daylight} => {Slight}	0.302	0.818	1.044
3	{Urban, Speed_30, Daylight} => {Slight}	0.281	0.817	1.043
4	{Urban, Dry_road, Fine without high winds, Daylight} => {Slight}	0.282	0.815	1.040
5	{Urban, Fine without high winds, Daylight} => {Slight}	0.317	0.815	1.040

Clustering: Identifying accidents in Kingston upon Hull, Humberside, East riding of Yorkshire etc.

In our study of the Identifying accidents in Kingston upon Hull, Humberside, East riding of Yorkshire etc, we explored different clustering methods such as DBSCAN and K-medoids, but ultimately chose to use the K-means algorithm with Euclidean distance to analyze the distribution of accidents based on longitude and latitude (Li et al., 2017). After using the elbow

method to determine the optimal number of clusters, we found that there were a total of 5 clusters in the region (Fig. 12). Our results showed that the majority of accidents occurred in and around major cities like Hull, Scunthorpe, and Bridlington (Fig. 13).

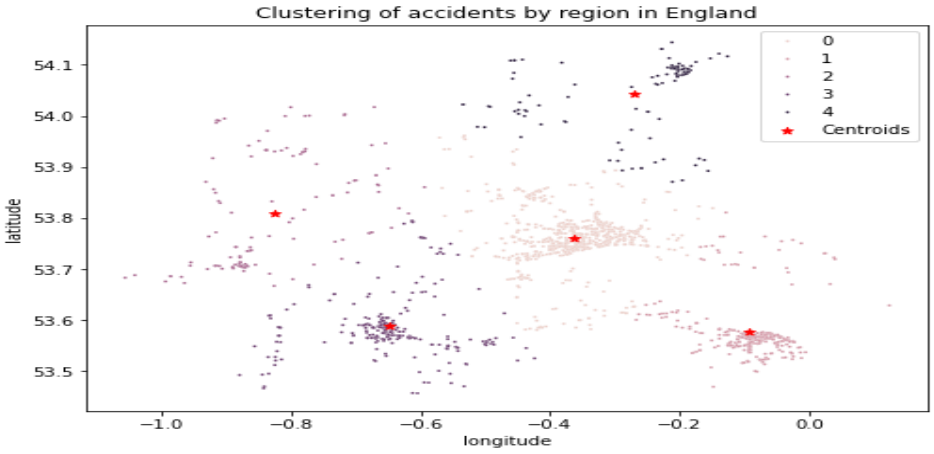

Fig. 11: Accident Region in UK

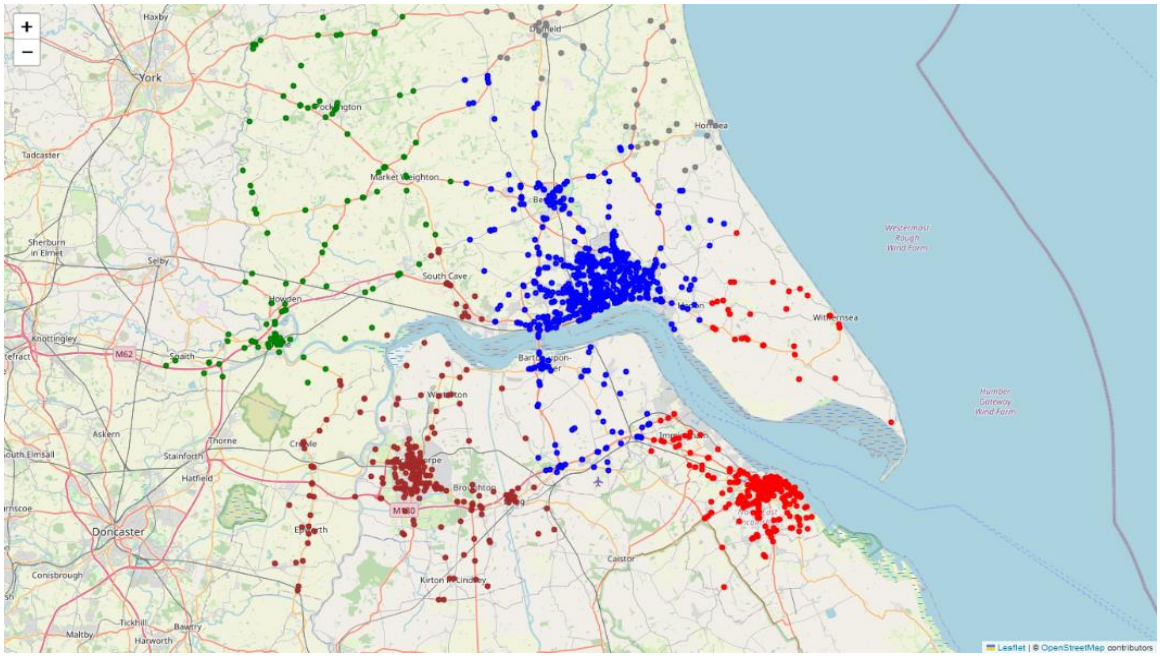

Fig. 12: Clustering with Accident Region in UK

Detecting Outlier

Outlier detection techniques, including IQR, Grubbs test, Isolation Forest, and Local Outlier Factor, were utilized to analyze specific features of accident data. The results showed that outliers were present throughout the dataset, with a smaller concentration near the London area (Fig. 13). This localized clustering suggests that certain factors or circumstances may have contributed to these anomalous incidents. Further analysis in the Humberside region (Fig.14) revealed that the outliers were not extreme values, indicating they were not caused by data entry errors or measurement inaccuracies. These outliers may represent unique or uncommon situations within the city, and as such, they were kept in the dataset.

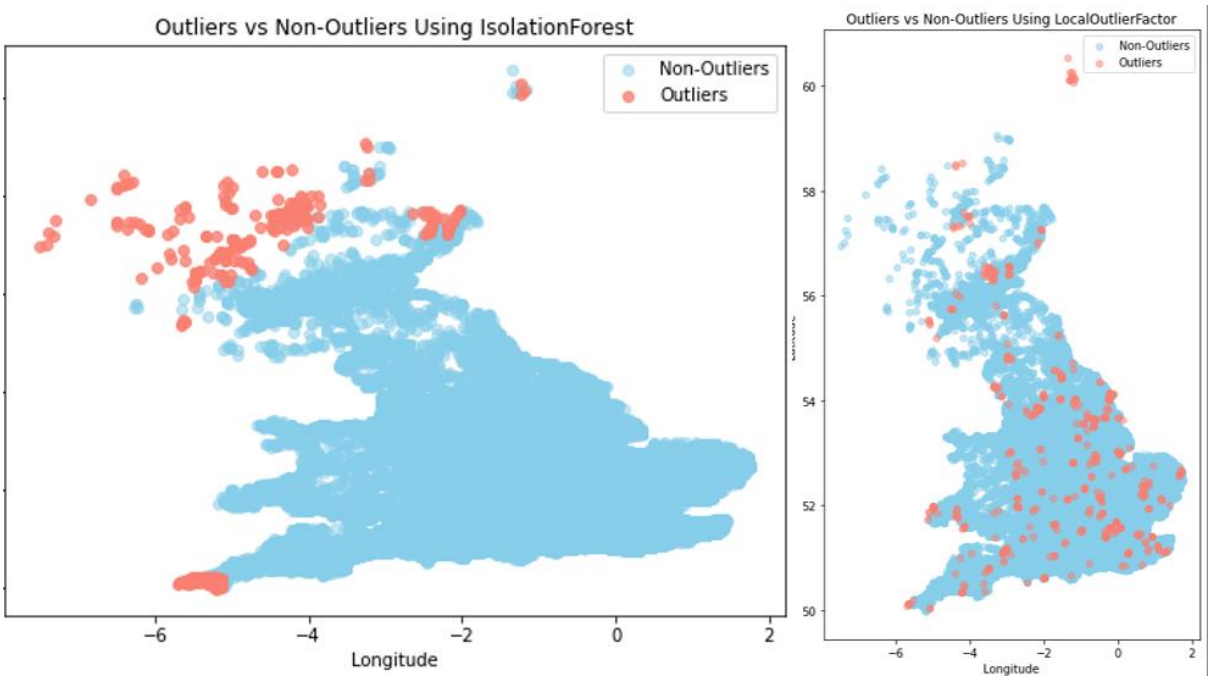

Fig. 13: Outlier Distribution

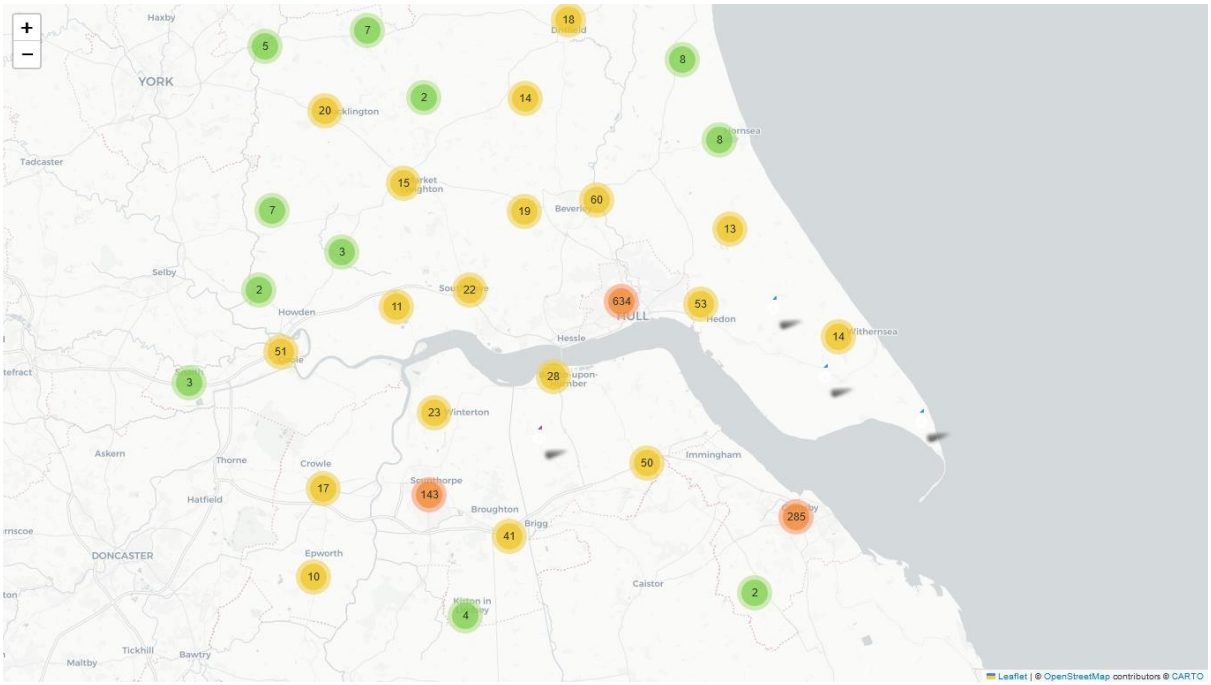

Fig. 14: Humberside Outliers

Prediction

To develop an accurate predictive model, it is crucial to utilize techniques such as feature balancing and normalization, specifically under-sampling and standard scaling. These methods were applied by Haynes et al. (2019) when selecting relevant features using Random Forests (RFs) based on their importance indices. The visual representation of feature importance rankings from the RFs can be seen in Fig. 15. In the analysis of accident severity prediction in the UK for 2020, the top 11 features were selected and SMOTE (Synthetic Minority Over-sampling Technique) was used to handle imbalanced data. Four classifiers were applied: Random Forest, Decision Tree, Logistic Regression, and Gradient Boosting Classifier. The initial results indicated that the Random Forest classifier performed the best among these models, achieving an accuracy of 0.9248, precision of 0.9242, recall of 0.9242, and an F1 score of 0.9242. In comparison, the Gradient Boosting Classifier had an accuracy of 0.6435, the Decision Tree had 0.8509, and the Logistic Regression had 0.5585.

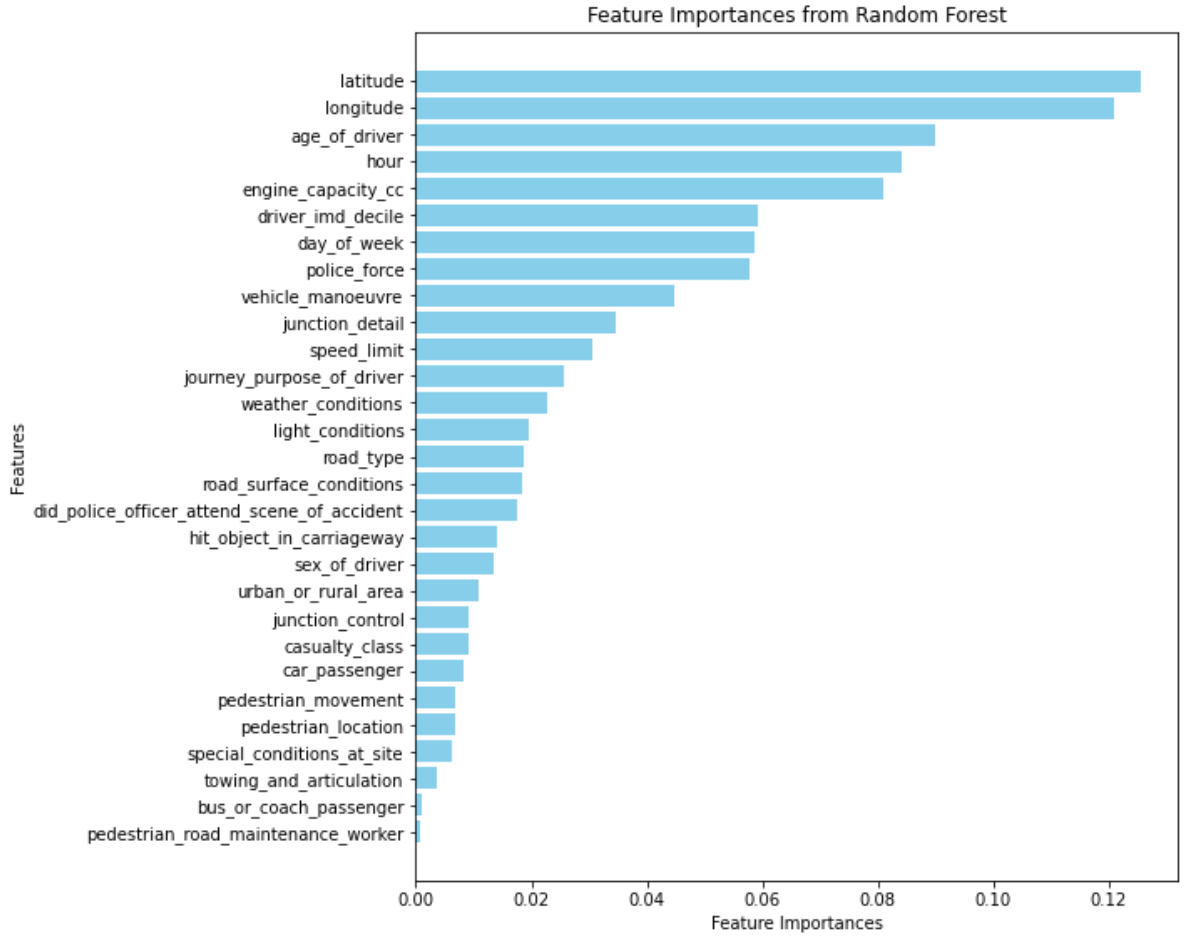

Fig. 15: Features Importances

After hyperparameter tuning the Random Forest classifier achieved an accuracy of 0.9243, with precision, recall, and F1 score all at 0.9242. The Decision Tree classifier resulted in an accuracy of 0.8458, with precision, recall, and F1 score at 0.8447 and 0.8449 respectively. Logistic Regression yielded an accuracy of 0.5695, with precision, recall, and F1 score all around 0.5642. Finally, the Gradient Boosting Classifier showed an accuracy of 0.6544, with precision, recall, and F1 score at 0.6489 and 0.6488 respectively. These results confirm that the Random Forest classifier consistently outperforms the other models in predicting accident severity, both before and after hyperparameter tuning (Fig. 16). The Decision Tree also shows relatively high performance, though not as high as the Random Forest. Logistic Regression and Gradient

Classification Report feature inportance Random forest classifier:

	precision	recall	f1-score	support
1	0.99	0.99	0.99	30507
2	0.90	0.88	0.89	30501
3	0.89	0.90	0.90	30453
accuracy			0.92	91461
macro avg	0.92	0.92	0.92	91461
weighted avg	0.92	0.92	0.92	91461

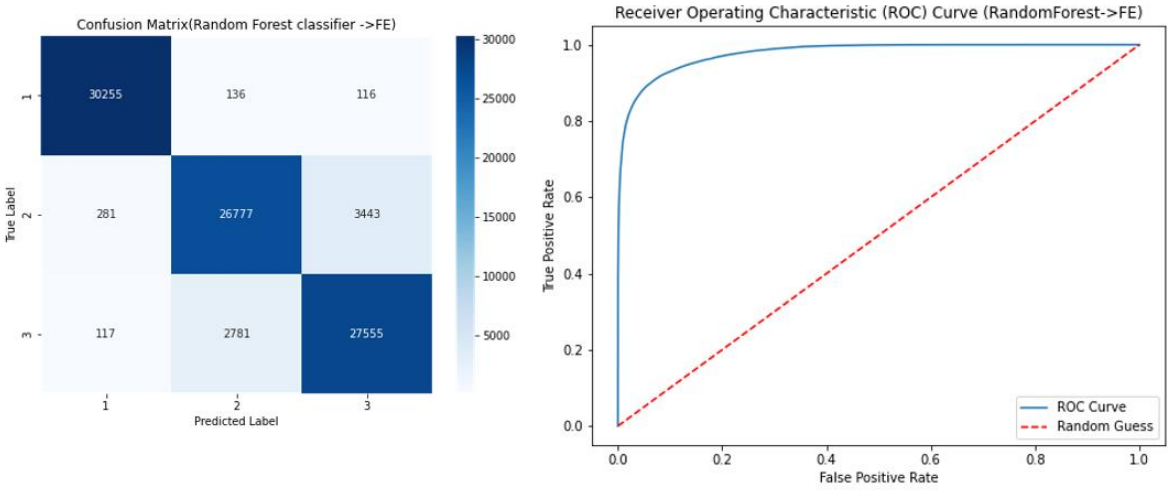

Fig. 16: Random Forest classification report, confusion matric, and ROC Curve

Boosting Classifier perform less effectively, with Logistic Regression being the least accurate among the four models.

Even with stacking, the Random Forest model outperformed all other classification models (as seen in Figure 17). In conclusion, this model has strong predictive capabilities for identifying fatal accidents with a high level of accuracy.

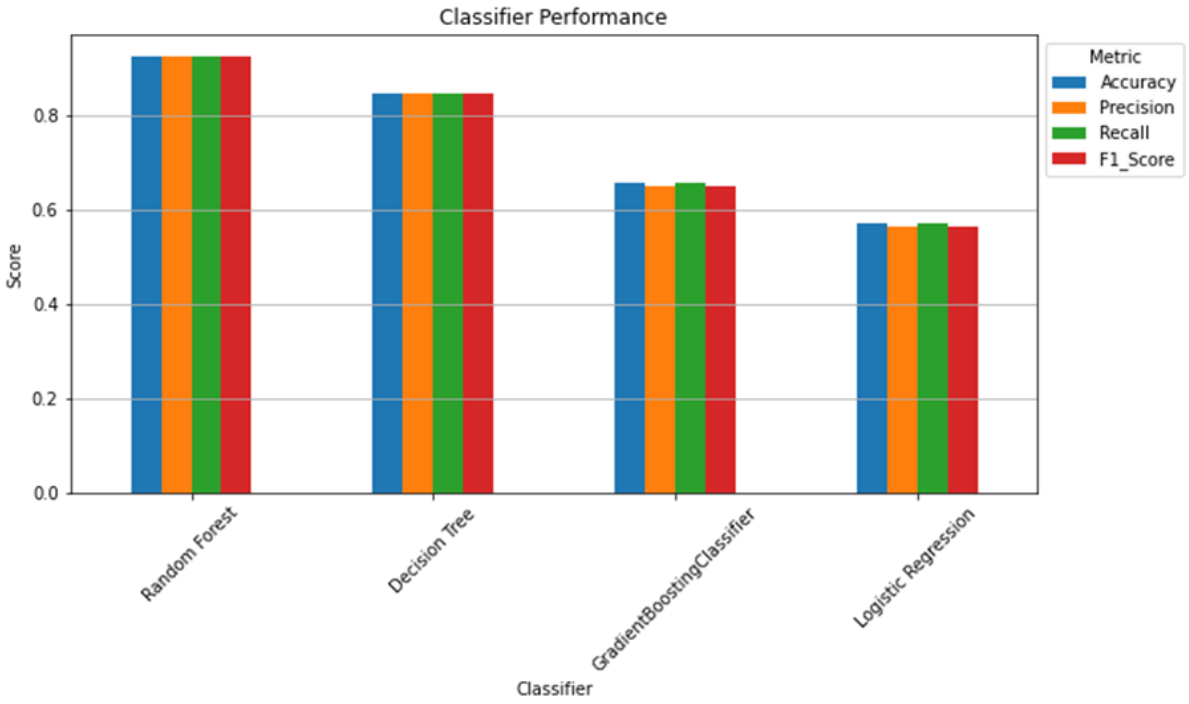

Fig. 18: Model performance metric of different classification algorithms

Recommendations

Based on the comprehensive data analysis and prediction, the following recommendations are proposed: 1. Optimize Urban Roads: Instead of expanding busy roads during rush hours or converting them to dual carriageways, consider implementing alternative solutions such as promoting public transportation or implementing traffic calming measures to reduce accidents. 2. Encourage Safer Modes of Transportation: Instead of solely focusing on promoting electric motorcycles, consider promoting a variety of safer modes of transportation such as bicycles or electric scooters. 3. Balance Mobility and Pedestrian Safety: Instead of enacting strict mobility scooter rules, find a balance between mobility and pedestrian safety by implementing designated scooter lanes and speed limits in walkways. 4. Improve Traffic Control: Instead of solely relying on traffic officers to ensure adherence to traffic rules, consider implementing technological solutions such as traffic cameras or sensors to enhance 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.