Analysis of 2020 UK Accident Data

Introduction

Road traffic accidents continue to be a major cause of injuries and fatalities, thus becoming an urgent concern for people all over the world (Gururaj, 2008). Gopalakrishnan (2012) mentioned that numerous lives are impacted by these accidents every year, hence it is critical to address the underlying causes and put in place efficient preventative measures. Ahmed et al. (2023) corroborated this by estimating 1.35 million deaths globally linked to road accidents. Consequently, governments and organizations reduce these risks through infrastructure improvements, traffic laws, and public awareness programs. (Zegeer & Bushell 2012).

The accident database provides valuable insights into road accidents, offering a comprehensive collection of accident-related information. The aim is to address crucial challenges in road safety and accident prevention for the year 2020 by analysing various factors, identify patterns and correlations that contribute to accident severity and build accurate Machine Learning models that can predict the likelihood of accidents and its severity.

Data Analysis & Insights

Outlier Detection and Data Cleaning.

To detect outliers, three common techniques were used thus: the Z-score and Local Outlier Factor (LOF) methods and Isolation Forest. The Z-score measures how far a data point is from the mean in terms of standard deviations (Schober et al., 2021). Data points with a Z-score beyond the threshold of 3 were considered outliers resulting in 76724 rows in 74 columns across the database. On the other hand, Xu et al. (2022) elaborated on how LOF calculates the local density deviation of a data point with respect to its neighbours. With LOF, there were 20195 rows across 74 columns with points with a significantly lower density than their neighbours. Lastly Kumar et al., (2022) identified outliers as having significantly different characteristics from the dataset.

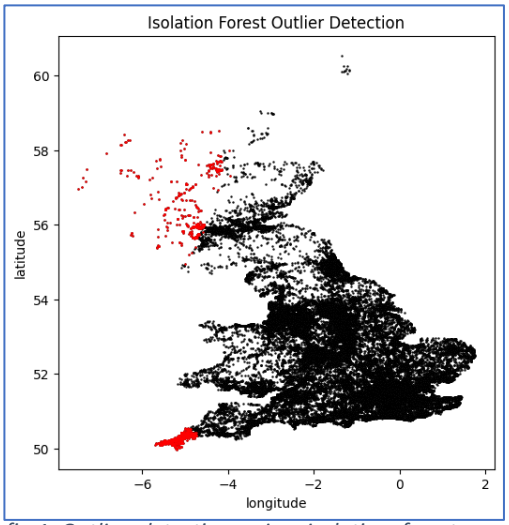

fig.1_Outlier_detection_using_isolation_forest

The data was particularly laced with some white spaces and some errors. To correct these, 14 whitespaces in location_easting_osgr, location_northing_osgr, longitude and latitude columns were replaced with the median value of each column. Subsequently, columns containing -1, 99, 999 entries were identified; categorical columns were replaced with the mode of the respective column whilst the continuous columns were replaced using the median value of the respective column. Compared to the median, the mean is more susceptible to outliers and more affected by the distribution of values hence, its usage in replacing the errors and missing values (Thompson, 2009). Also, based on GOV.UK, only people aged 17 and above are allowed to drive hence 2519 ages below 17 were corrected with the median age for drivers. Finally, the mode was used to replace nominal values to minimise the potential bias by keeping the data centred on the most prevalent category (Zhang, 2016).

Accident counts per Hour.

As shown in fig.2, accidents are most likely to occur from 5pm to 6pm, which records 7,813 accidents, followed closely by 4pm to 5pm with 7,381 accidents. Conversely, 4am to 5am recorded the least number of accidents. Similarly, Meyyappan et al. (2018) concluded their research stating that road accidents are likely to occur between 7am - 10am and 5pm - 9pm, when there is usually heavy traffic. Interestingly, most accidents beyond the median mark of 3,826 accidents occur between 8am and 7pm, the hours encompassing the typical working day when traffic volume is generally higher.

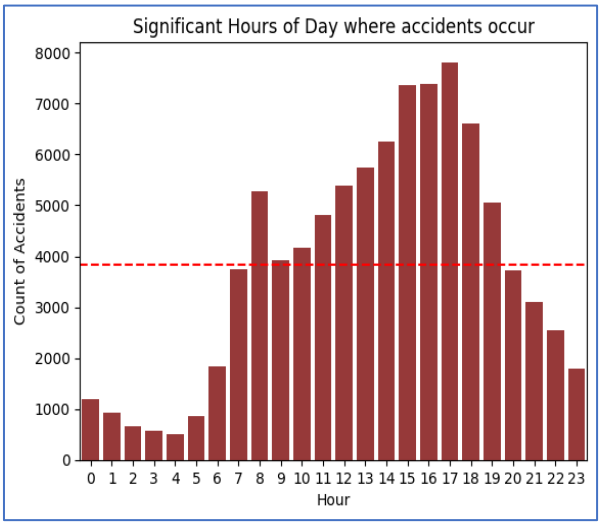

fig.2_Accident_counts_per_hour

Accident counts per Day.

Unlike Farmer and Williams, (2005) who argued that accidents are most likely to occur on weekends, analysis of the accident database revealed that accidents on Fridays account for 16.3% of total accidents followed closely by Thursdays with 15.4% of the accidents, while Saturdays and Sundays both have around 13.5% and 11.3% of the accidents, respectively. The significant number of accidents on Fridays can be attributed to increased vehicular and pedestrian traffic as people often engage in various social and recreational activities (Damsere-Derry et al., 2010). As a result, roadways tend to be more congested on Fridays, leading to a higher likelihood of accidents (Cabrera-Arnau, Prieto Curiel, and Bishop., 2020)

Table1._accident_counts_per_day

	Days	Count of Accidents	Percentage
0	Friday	14889	16.3
1	Thursday	14056	15.4
2	Wednesday	13564	14.9
3	Tuesday	13267	14.5
4	Monday	12772	14.0
5	Saturday	12336	13.5
6	Sunday	10315	11.3

Motorcycle accident counts per Hour.

Analysing motorcycle accidents involved a focus on 3 groups of motorcycles; 125cc and below, between 125cc and 500cc and lastly motorcycles over 500c. The hours with the significant occurrences of motorcycle accidents are between 3pm and 6pm, with the peak being observed at 5pm with 1401 accidents. This pattern suggests that late afternoons and early evenings pose a higher risk for motorcycle accidents (Clarke et al., 2004). Conversely, the early morning hours, from 2am to 5am, show relatively lower accident counts, indicating that these times might be comparatively safer for motorcycle riders. Motorcycles with 125cc-500cc and over 500cc had 5pm as their significant accident hours whilst motorcycles 125cc and below recorded more accidents at 6pm.

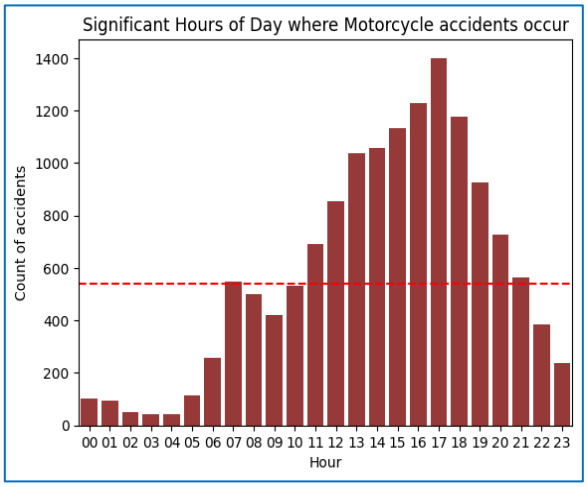

fig.3_Motorcycle_accidents_per_hour

Motorcycle accident counts per Day.

Similarly, Fridays are associated with the highest number of motorcycle accidents (Rzeznikiewiz et al., 2012). There were 2,308 motorcycle accidents on Fridays followed by 2172 on Thursdays. However, Sunday and Monday saw the least motorcycle accidents with 1841 and 1868 respectively. Fig.4 shows the distribution of motorcycle accidents per day. Narrowing down, lower engine capacity motorcycles contributed to the most accidents on Fridays followed by Thursday with 1474 and 1389 accidents respectively. Those with engine capacity between 125cc and 500cc were involved in 267 and 257 accidents on Friday and Thursdays respectively. Even though, the first two groups of motorcycles recorded the least number of accidents on Sundays, the larger engine sized motorcycle with 500cc and above were involved in 675 accidents on Sunday and 567 accidents on Fridays with its least being on Mondays. The high number of accidents associated with 500cc and above motorcycles on Sundays could be connected to lesser traffic on Sundays (Lonati et al., 2006). As a result, these riders tend to speed resulting in accidents (Clarke et al., 2004).

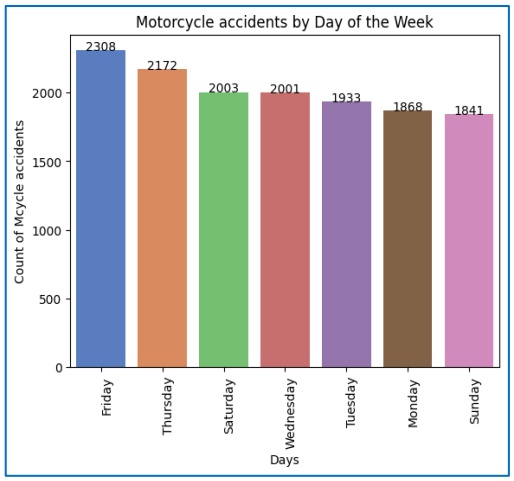

fig.4_Motorcycle_accidents_per_day

Pedestrian involved in accidents per Hour.

Schieber and Vegega (2002) stated that pedestrians are particularly involved in accidents between 3pm and 6pm. A substantial portion of pedestrian accidents occur during these three hours, and as the night wears on, fewer incidences occur. From our data, 5415 incidents occurred with pedestrians from 3pm to 6pm making up 37% of pedestrian accidents in the year 2020. The highest hour with most accidents was 3pm with 1672 accidents. Conversely, the late-night hours, from midnight to 5am, exhibits relatively lower pedestrian accident counts with 606 incidents at 4%. 4pm recorded the lowest hour for pedestrian accidents with 50 accidents.

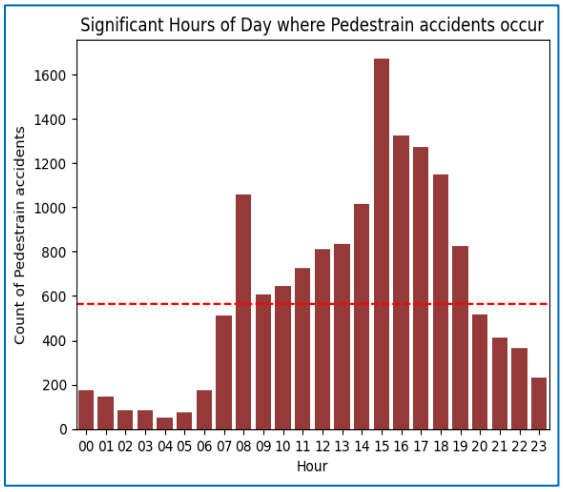

fig.5_pedestrian_accidents_per_hour

Pedestrian involved in accidents per Day.

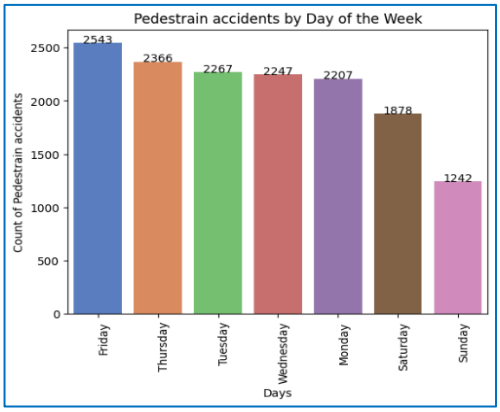

fig.6_pedestria_accidents_per_day

Fig.6 illustrates the count of pedestrian accidents recorded on different days of the week. Similarly, it is evident that Friday has the highest number of pedestrian accidents with a count of 2543, making it the most accident-prone day for pedestrians. Thursday follows closely with 2366 accidents, indicating a significant risk for pedestrians during mid-week. Research by Rosenbloom (2009) supported these conclusions, indicating that pedestrian accidents are most likely to occur on Fridays. Interestingly, weekends seem to have relatively lower pedestrian accident counts with Sunday recording the lowest incidence (Malin et al., 2020).

Association mining based on Accident severity.

Association mining techniques were used to analyse the UK traffic accident data which revealed unique linkages and patterns associated with accidents. By employing the Apriori algorithm and setting a minimum support of 0.2 and a confidence threshold of 0.7, a comprehensive set of association rules were generated. However, it's important to note that due to the strict criteria, the focus of the analysis primarily centred on accidents severity 3 (denoting slight accidents), which had a notably high frequency within the dataset. Hence, 434 association rules were identified as illustrated in fig.7.

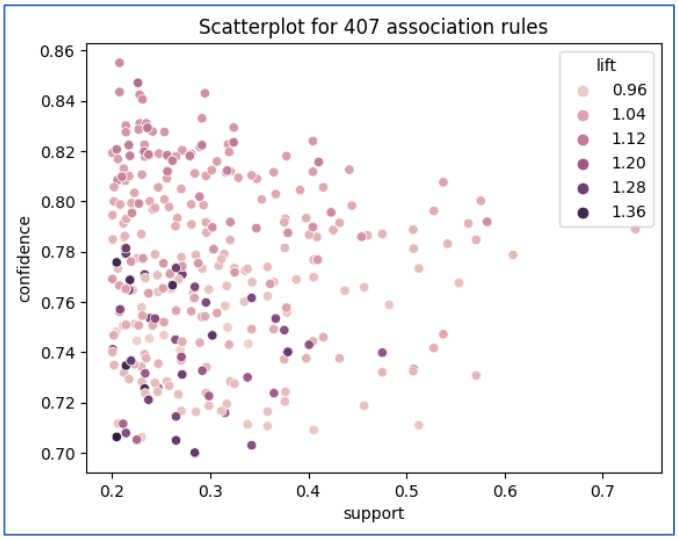

fig.7_Scatterplot_showing_association_rules_based_on_accident_severity.

To analyse these rules, the association pattern was firstly sorted by the Support which describes the frequency with which accident occurs. It identified pedestrian_location_0 = unknown location, weather_conditions_1 = fine, Casualty_class_1 = Driver/Rider, Vehicle_type_9 = cars, light_conditions_1 = daylight, road_type_6 = single carriageway, road_surface_conditions_1 = dry, association between casualty_class_1 and pedestrian_location_0, pedestrian_location_0 and weather_conditions_1 etc has high frequency rules linking to accident_severity_3. Similarly, Wahab and Jiang (2019) in their research concluded on the possible effects that vehicle type, road and weather conditions have on accident severity.

Further analysis based on confidence, highlighted the strength of relationships between antecedents and consequents. Notably, vehicle type, light conditions, speed limit, and casualty class emerged as key contributors to accident severity. The high confidence levels suggest a reliable predictive power of these variables in determining accident severity. Shaheed et al. (2011) corroborated the effect of lighting conditions on the causes of road accidents in their study.

Lastly, the lift-based analysis emphasized the importance of considering the joint occurrence of attributes (Edirisinghe & Munson, 2023). Certain combinations, such as specific road surface conditions, weather conditions, and pedestrian locations, demonstrated higher-than-expected lift values, indicating a non-random relationship between these factors and accident severity.

These associations are mined to help governments and local authorities better understand the underlying causes of accidents and its severity Fig 8,9 and 10 below depicts heatmaps for the above.

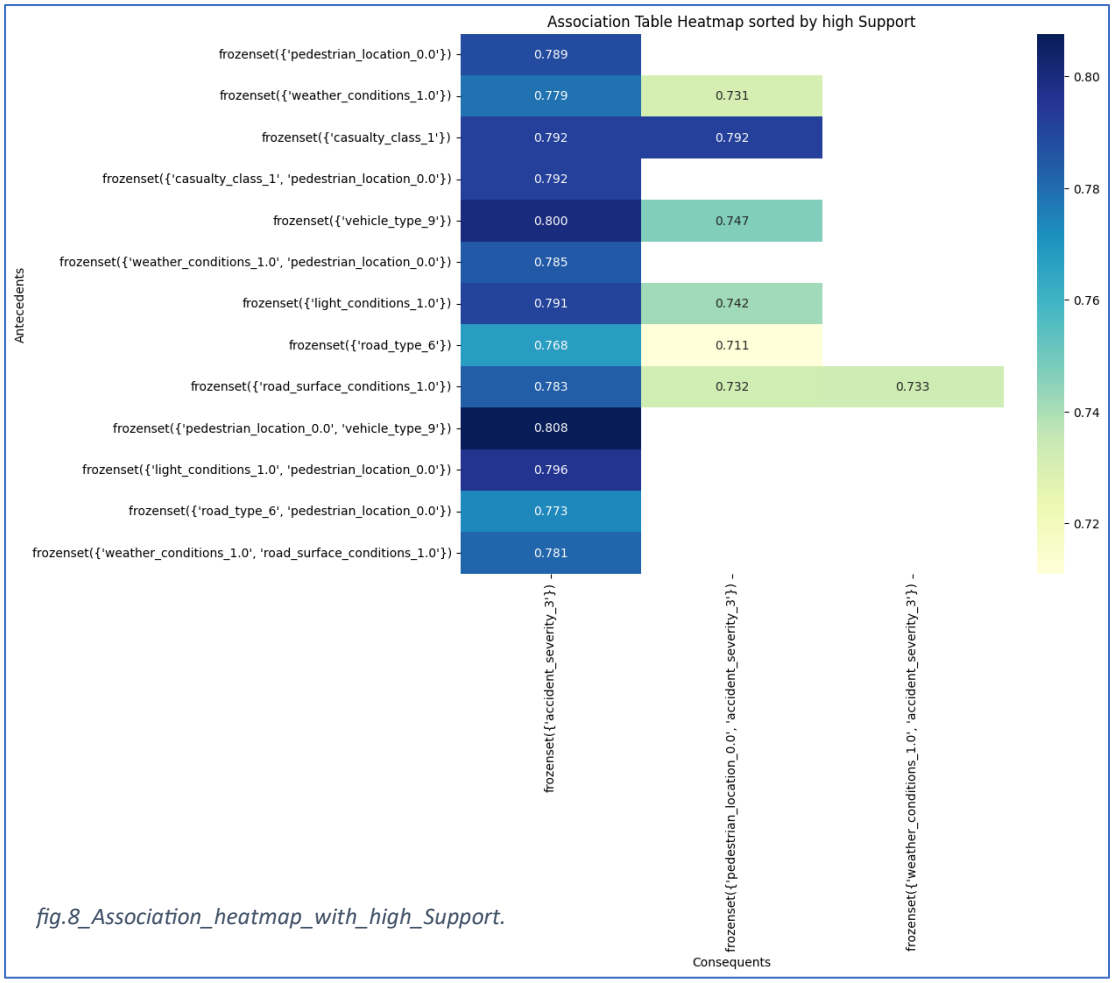

fig.8_Association_heatmap_with_high_Support.

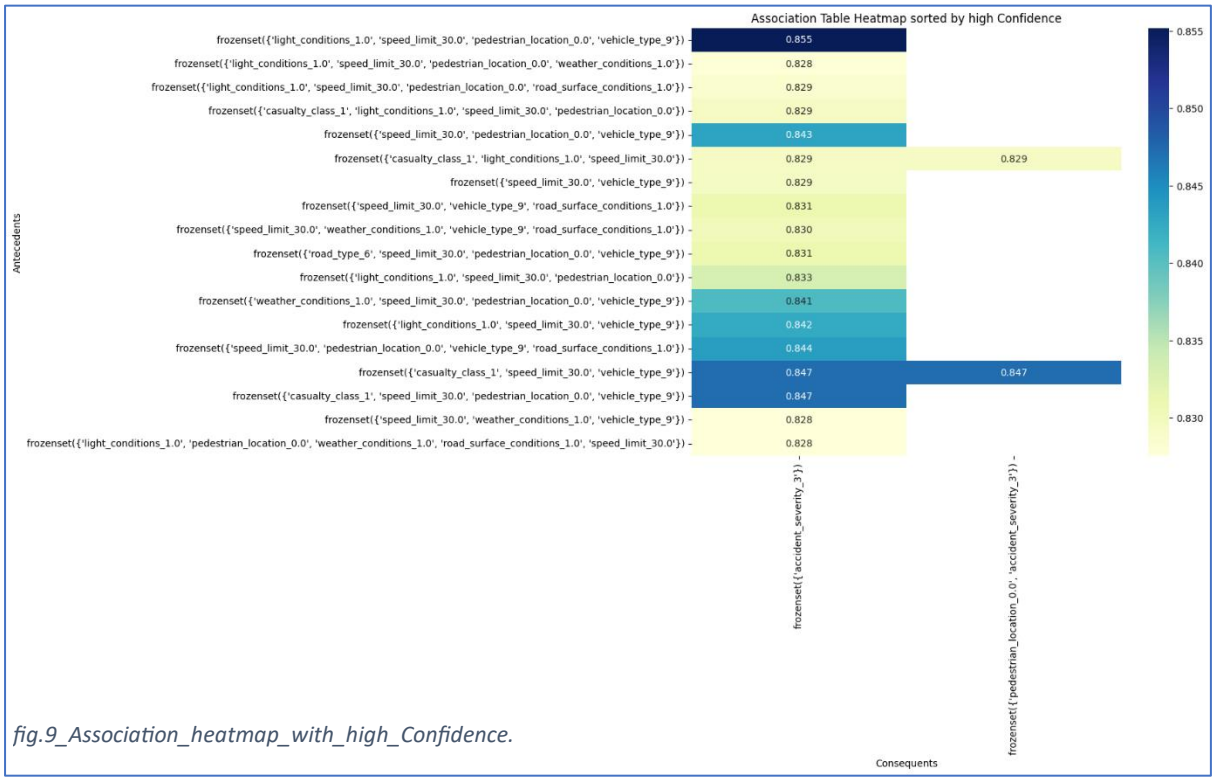

fig.9_Association_heatmap_with_high_Confidence.

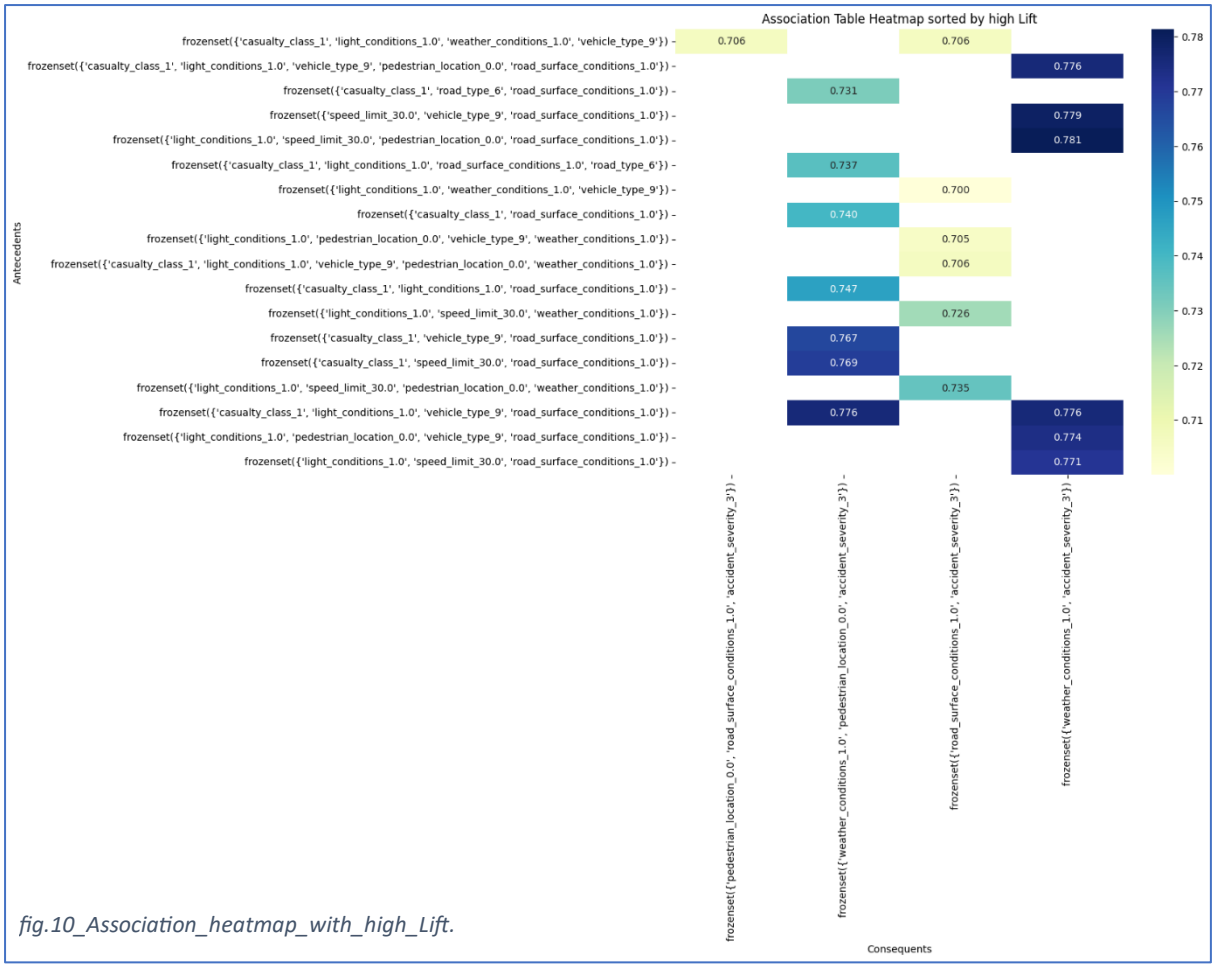

fig.10_Association_heatmap_with_high_Lift.

Focused Clustering of Accidents in Humberside Region.

A K-means clustering algorithm was applied to the geographical coordinates (longitude and latitude) of accident locations in Humberside. The algorithm identified four distinct clusters that are visually well-represented on a scatterplot namely Kingston upon Hull (Hull), East Riding of Yorkshire (ERY), North Lincolnshire (NL), and Northeast Lincolnshire (NEL).

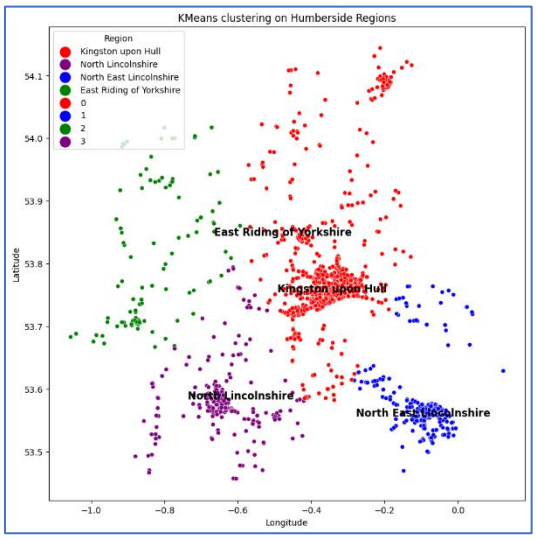

fig.11_K-Means_clustering_on_Humberside.

With reference to Hull, accidents predominantly occurred in densely populated urban areas, characterized by higher traffic volumes and mostly on single carriage roads. However out of 569 accidents in the year 2020, only 1.2% resulted in death. The story is same with NEL, which also shows a lot more accidents in the urban areas and on single carriage roads. However, ERY and NL, recorded more accidents in rural areas and smaller towns, where accidents were more sporadic. Interestingly, across all regions, speed limit 70 recorded fewer accidents than the lower limit tiers except for speed limit 50. This is corroborated by the number of accidents on single carriage roads and dual carriage roads where the speed limits are much higher. Lastly, 71% and 85% of all accidents within Humberside occurred during the day with fine weather.

Table_2_Regional_clustering_of_accident_factors

	accident_severity			urban_or_rural_area		speed_limit						road_type					
	1	2	3	1	2	20	30	40	50	60	70	1	2	3	6	7	9
East Riding of Yorkshire	12	95	381	163	325	34	205	45	28	155	21	31	0	38	412	6	1
Kingston upon Hull	7	96	466	556	13	22	491	52	1	0	3	56	0	121	386	6	0
North East Lincolnshire	2	61	239	242	60	11	238	11	16	14	12	33	1	32	235	1	0
North Lincolnshire	11	65	228	127	177	3	154	30	17	69	31	20	0	50	230	4	0
	32	**317**	**1314**	**1088**	**575**	**70**	**1088**	**138**	**62**	**238**	**67**	**140**	**1**	**241**	**1263**	**17**	**1**

Table_3_cont._Regional_clustering_of_accident_factors

	weather_conditions								light_conditions				
	1	2	3	4	5	7	8	9	1	4	5	6	7
East Riding of Yorkshire	417	42	0	15	5	2	6	1	354	69	3	60	2
Kingston upon Hull	484	47	1	9	7	3	15	3	401	159	5	1	3
North East Lincolnshire	254	25	0	10	1	2	8	2	208	79	5	10	0
North Lincolnshire	253	24	0	17	5	1	4	0	219	52	1	32	0
	1408	**138**	**1**	**51**	**18**	**8**	**33**	**6**	**1182**	**359**	**14**	**103**	**5**

Predicting Accident Severities using ML:

In the context of predicting accident severity using machine learning, several classifiers were employed to analyse and predict the severity of accidents thus fatal or non-fatal accidents. The classifiers used were Gaussian Naive Bayes, Logistic Regression, Decision Tree, Support Vector Classifier (SVC), Gradient Boosting Classifier, XGBoost Classifier, and Random Forest Classifier. Both unclean and cleaned data were used in training these classifiers.

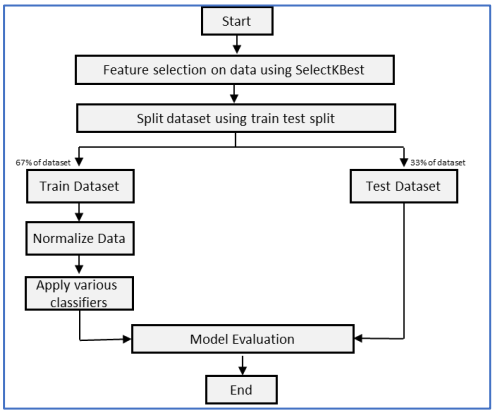

fig.12_Model_architecture

Being a binary classification problem, the target variable 'accident_severity' was highlighted and balanced, feature extraction carried out with SelectKBest tool. The datasets were split using train test split with test size of 33%. Training sets was normalised using StandardScaler. Subsequently the various classifiers were executed, and model evaluated.

The models were evaluated using accuracy, Weighted precision, recall, and F1-score, true positive (TP) rate, The false positive (FP) rate, Correctly Classified Instances (CCI), Incorrectly Classified Instances (ICI) (Kumeda et al., 2019). As expected, the unclean dataset showed a relatively poor performance as compared to the cleaned dataset (Wu, 2021). Model accuracy hovered from 61% to 72% as indicated in table 4 below. Conversely, model accuracy for cleaned dataset ranged from 0.86 to 0.94. showcasing the models' ability to correctly classify accident severity. Weighted precision, recall, and F1-score exhibited strong consistency across classifiers, with values around 0.86 to 0.94. These metrics indicate that the models can effectively predict both fatal and non-fatal classes.

Furthermore, TP rate and CCI were relatively higher whilst the FP rate and the ICI was relatively low indicating incorrectly predicted fatal cases.

Random Forest Classifier led the classifiers with an accuracy of 0.94 and scored highly on all other measures. Both Gradient Boosting and XGBoost had outstanding accuracy scores of 0.92 and 0.93, respectively. Overall, these findings imply that ensemble-based classifiers are effective at estimating accident severity based on the dataset.

Table_4._Model_performance_with_unclean_dataset

Classifier	Accuracy	Weighted Prescision	Weighted Recall	Weighted F1-Score	TP Rate	FP Rate	CCI	ICI
DecisionTreeClassifier	0.70	0.71	0.70	0.70	0.73	0.32	646	273
GaussianNB	0.61	0.64	0.61	0.58	0.71	0.43	556	363
RandomForestClassifier	0.69	0.69	0.69	0.69	0.70	0.32	632	287
SVC	0.72	0.72	0.72	0.72	0.72	0.28	663	256
GradientBoostingClassifier	0.72	0.72	0.72	0.72	0.73	0.29	659	260
XGBClassifier	0.69	0.69	0.69	0.69	0.70	0.31	636	283
LogisticRegression	0.70	0.70	0.70	0.70	0.70	0.30	644	275

Table_5._Model_performance_with_clean_dataset

Classifier	Accuracy	Weighted Prescision	Weighted Recall	Weighted F1-Score	TP Rate	FP Rate	CCI	ICI
GaussianNB	0.86	0.86	0.86	0.86	0.86	0.15	2077	347
LogisticRegression	0.88	0.88	0.88	0.88	0.87	0.11	2133	291
DecisionTreeClassifier	0.89	0.89	0.89	0.89	0.90	0.12	2166	258
SVC	0.90	0.90	0.90	0.90	0.90	0.09	2187	237
GradientBoostingClassifier	0.92	0.92	0.92	0.92	0.92	0.09	2223	201
XGBClassifier	0.93	0.93	0.93	0.93	0.93	0.08	2244	180
RandomForestClassifier	0.94	0.94	0.94	0.94	0.94	0.07	2267	157

Recommendations

Based on the analysis of the accident data, here are the top recommendations for government agencies to improve safety:

a) Enhance Road Infrastructure and Maintenance: Allocate resources to identify and prioritize high-risk road sections for infrastructure upgrades such as expansion of single carriage roads to dual carriage roads. Example, policies like speed limits, rumble strips etc need to be

instituted in areas where day light accident occurrences are linked to driver/rider, dry weather conditions.

b) Promote Driver Education and Awareness: Launch targeted campaigns to educate drivers about safe driving practices, safe urban driving, defensive driving, speeding and distracted driving.

c) Enforce Stricter Traffic Regulations: Strengthen enforcement of traffic regulations, especially in high-risk areas, by increasing the presence of law enforcement officers and deploying speed cameras. Also, impose stricter penalties for traffic violations, including fines, license suspension, and mandatory driving courses for repeat offenders.

d) Implement Intelligent Traffic Management Systems: Invest in intelligent traffic management systems that leverage data analytics and real-time monitoring to identify congestion, accidents, and road hazards promptly. Use predictive analytics to anticipate traffic patterns, wet roads for salt-spreading and mitigate potential issues before they escalate (Kuttesch, 2004).

By implementing these recommendations, government agencies can proactively enhance road safety, reduce the frequency of accidents, and ultimately save lives.

References:

Ahmed, S.K., Mohammed, M.G., Abdulqadir, S.O., El-Kader, R.G.A., El-Shall, N.A., Chandran, D., Rehman, M.E.U. and Dhama, K., 2023. Road traffic accidental injuries and deaths: A neglected global health issue. *Health science reports*, *6*(5), p.e1240.

Cabrera-Arnau, C., Prieto Curiel, R. and Bishop, S.R., 2020. Uncovering the behaviour of road accidents in urban areas. *Royal Society open science*, *7*(4), p.191739.

Clarke, D.D., Ward, P., Bartle, C. and Truman, W., 2004. In-depth study of motorcycle accidents. *Road Safety Research Rep*, *54*.

Clarke, D.D., Ward, P., Bartle, C. and Truman, W., 2004. In-depth study of motorcycle accidents. *Road Safety Research Rep*, *54*.

Damsere-Derry, J., Ebel, B.E., Mock, C.N., Afukaar, F. and Donkor, P., 2010. Pedestrians' injury patterns in Ghana. *Accident Analysis & Prevention*, *42*(4), pp.1080-1088.

GOV.UK *Driving lessons and learning to drive*
Available online:
Driving lessons and learning to drive: Overview - GOV.UK (www.gov.uk)
[Accessed: 10-08-2023]

Edirisinghe, G.S. and Munson, C.L., 2023. Strategic rearrangement of retail shelf space allocations: Using data insights to encourage impulse buying. *Expert Systems with Applications*, *216*, p.119442.

Farmer, C.M. and Williams, A.F., 2005. Temporal factors in motor vehicle crash deaths. *Injury Prevention*, *11*(1), pp.18-23.

Gopalakrishnan, S., 2012. A public health perspective of road traffic accidents. *Journal of family medicine and primary care*, *1*(2), p.144.

Kumar, S.G., Corrado, S.J., Puranik, T.G. and Mavris, D.N., 2022. Application of isolation forest for detection of energy anomalies in ADS-B trajectory data. In *AIAA SCITECH 2022 Forum* (p. 2441).

Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A. and Assefa, M., 2019, June. Classification of road traffic accident data using machine learning algorithms. In *2019 IEEE 11th international conference on communication software and networks (ICCSN)* (pp. 682-687). IEEE.

Kuttesch, J.S., 2004. *Quantifying the relationship between skid resistance and wet weather accidents for Virginia data* (Doctoral dissertation, Virginia Tech).

Lonati, G., Giugliano, M. and Cernuschi, S., 2006. The role of traffic emissions from weekends' and weekdays' fine PM data in Milan. *Atmospheric Environment*, *40*(31), pp.5998-6011.

Malin, F., Silla, A. and Mladenović, M.N., 2020. Prevalence and factors associated with pedestrian fatalities and serious injuries: case Finland. European transport research review, 12(1), pp.1-17.

Meyyappan, A., Subramani, P. and Kaliamoorthy, S., 2018. A comparative data analysis of 1835 road traffic accident victims. *Annals of maxillofacial surgery*, *8*(2), p.214.

Gururaj, G., 2008. Road traffic deaths, injuries and disabilities in India: current scenario. *National Medical Journal of India*, *21*(1), p.14.

Rosenbloom, T., 2009. Crossing at a red light: Behaviour of individuals and groups. *Transportation research part F: traffic psychology and behaviour*, *12*(5), pp.389-394.

Rzeznikiewiz, D., Tamim, H. and Macpherson, A.K., 2012. Risk of death in crashes on Ontario's highways. *BMC Public Health*, *12*(1), pp.1-7.

Schieber, R.A. and Vegega, M.E., 2002. Reducing childhood pedestrian injuries. *Inj Prev*, *8*(suppl 1), pp.i1-i10.

Schober, P., Mascha, E.J. and Vetter, T.R., 2021. Statistics from A (agreement) to Z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research. *Anesthesia & Analgesia*, *133*(6), pp.1633-1641.

Shaheed, M.S.B., Zhang, W., Gkritza, K. and Hans, Z., 2011, September. Differences in motorcycle conspicuity-related factors and motorcycle crash severities in daylight and dark conditions. In *3rd International Conference on Road Safety and Simulation. Indianapolis* (pp. 1-22).

Thompson, C.B., 2009. Descriptive data analysis. *Air medical journal*, *28*(2), pp.56-59.

Wahab, L. and Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. *PLoS one*, *14*(4), p.e0214966.

World Health Organization, 2015. *Global status report on road safety 2015*. World Health Organization.

Wu, X., Zheng, W., Xia, X. and Lo, D., 2021. Data quality matters: A case study on data label correctness for security bug report prediction. *IEEE Transactions on Software Engineering*, *48*(7), pp.2541-2556.

Xu, H., Zhang, L., Li, P. and Zhu, F., 2022. Outlier detection algorithm based on k-nearest neighbors-local outlier factor. *Journal of Algorithms & Computational Technology*, *16*, p.17483026221078111.

Zegeer, C.V. and Bushell, M., 2012. Pedestrian crash trends and potential countermeasures from around the world. *Accident Analysis & Prevention*, *44*(1), pp.3-11.

Zhang, Z., 2016. Missing data imputation: focusing on single imputation. *Annals of translational medicine*, *4*(1).