STUDENT NAME: BOBOYE DAMILARE ADEJUWON

STUDENT NUMBER: 202346817

MODULE: BIG DATA AND DATA MINING, 771762_A24_T2.

Big Data and Data Mining Methods for Road Accident Analysis and Prediction: A Case Study of Kingston upon Hull and Surrounding Regions

Abstract

This study comprises a comprehensive analysis of road traffic accident data using big data and data mining techniques to uncover underlying patterns, predict accident occurrences, and inform safety strategies in Kingston Upon Hull and surrounding areas. Extracting data from SQLite relational database, this report explores exploratory data analysis, association rule mining, cluster analysis, spatiotemporal analysis, social network analysis, and time series forecasting. Results revealed that accidents occur more frequently during peak hours on weekdays, particularly involving motorcyclists and pedestrians. The Apriori algorithm showed that drivers are the most frequently involved in slight accidents. Clustering techniques identified regional hotspots, while social network analysis exposed key structural influencers in communication patterns. Time series models, including SARIMA, ARIMA, and XGBoost, were applied to forecast accidents, with XGBoost outperforming statistical models across all selected regions. Based on the findings, the report recommends dynamic traffic control systems, targeted pedestrian awareness campaigns, and stricter vehicle licensing procedures.

INTRODUCTION

1.1 Background to the Study

This report analyses road traffic accident data from an SQLite relational database containing detailed information on accidents, vehicles, casualties, and Lower Layer Super Output Areas (LSOAs). Given urban regions' high traffic and pedestrian activity, effective analysis is essential to improving public safety. Insights gleaned from this study aim to support government efforts in developing policies that reduce the risk and frequency of future accidents.

ANALYSIS

2.1 Data Preparation and Cleaning:

To ensure proper data cleaning, I performed an in-depth Exploratory Data Analysis (EDA) using **Sweetviz** and **Pandas Profiling**. These libraries provided comprehensive reports highlighting data distributions, correlations, and missing values. A significant number of -1 and 9 entries were identified, which, according to the Road Safety Open Dataset Guide, represent missing values and unknown, respectively. These were addressed using the **Iterative Imputer** from the *scikit-learn* library—a machine learning technique that imitates unique missing data as a function of other features in a sequential manner (Prakash et al., 2024). Other missing values were imputed using the mean and median. Furthermore, labels from the road safety open dataset guide were assigned to their corresponding columns, ensuring robust analysis.

2.2 Temporal Analysis: When do Accidents Happen?

It is imperative to understand the significant hours, and days of the week accidents happen to improve safety measures. Also, it is beneficial to understand the engine capacity of vehicles and the casualty type and under what conditions accidents occur. For instance, as shown in Figure 1, data revealed that across a 24-hour period, accidents peak towards the evening. This pattern likely reflects increased traffic during the evening rush hour.

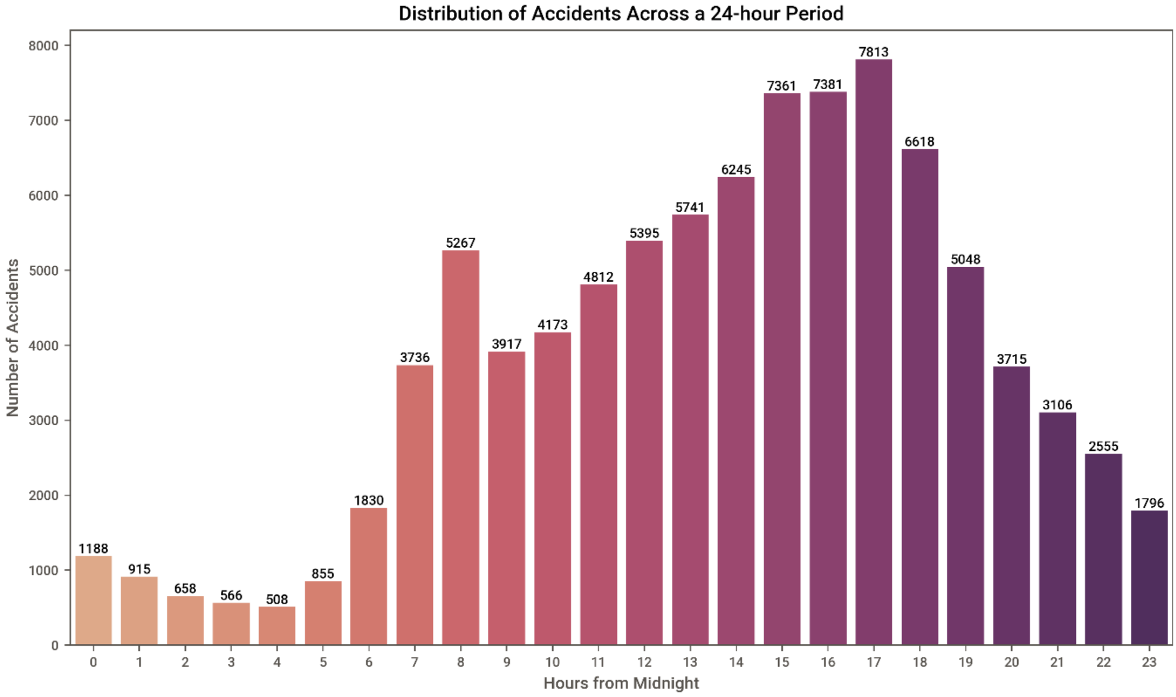

Figure 1: Accident Distribution by Hour of Day

Furthermore, the data indicates a higher frequency of accidents occurring during weekdays, as shown in Figure 2. This is reflective of increased commuter traffic and overall road activity during typical workdays.

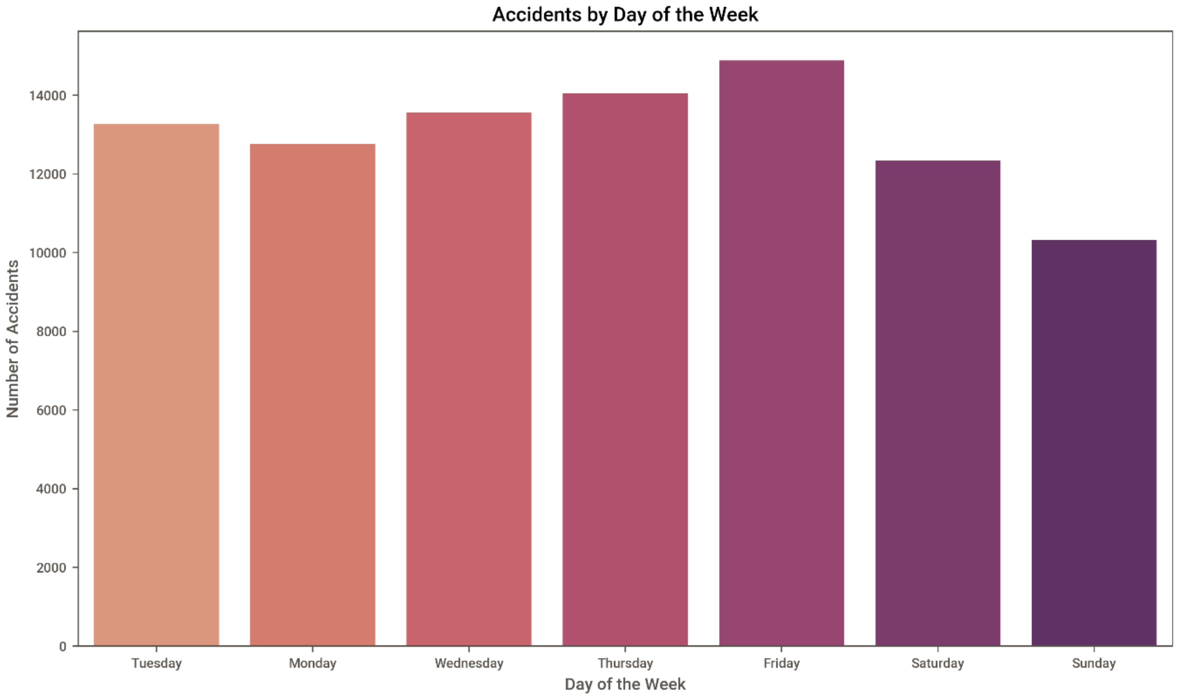

Figure 2: Accident Distribution by Day of Week

Furthermore, we analysed the engine capacity of motorcycles, particularly motorcycles with an engine capacity of 125cc and under, over 125cc and up to 500 cc, and over 500 cc. This measurement is in cubic centimeters and measures how much air and fuel the motorcycle engine can displace, which directly relates to the motorcycle's power and speed.

As shown in Figures 3, 4, and 5, accident trends differ across motorcycle engine capacities, with 125cc and under being more commonly involved in urban accidents during commuting hours on weekdays. According to the factsheet by the United Kingdom Government (2022), 13,604 motorcycle accidents occurred in 2020, resulting in either fatal, serious, or slight casualties.

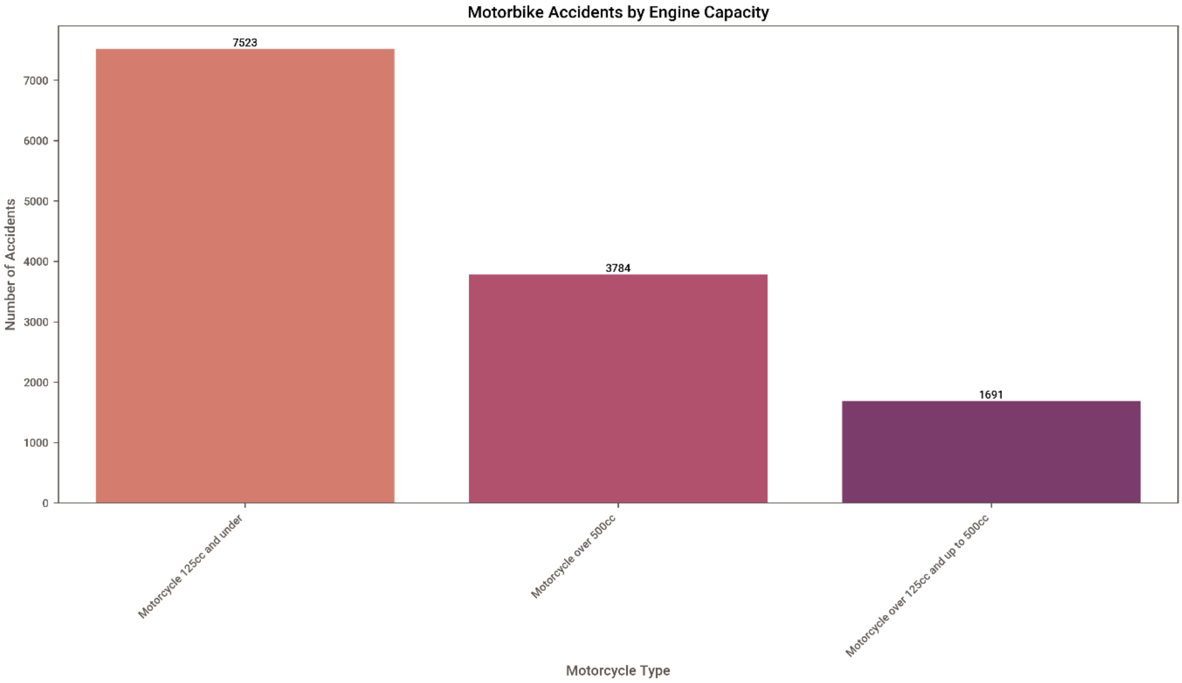

Figure 3: Motorbike Accidents by Engine Capacity

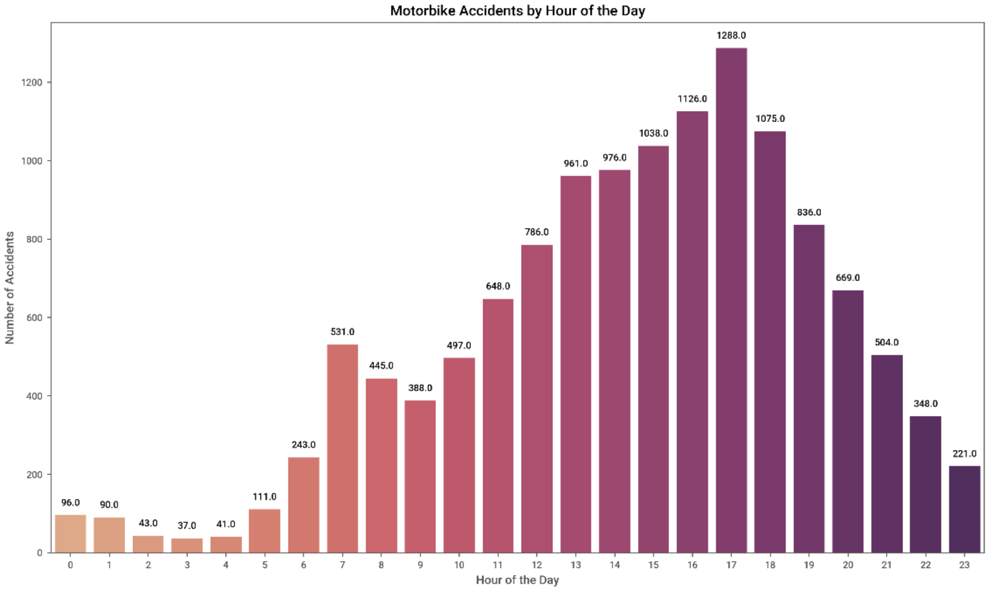

Figure 4: Motorbike Accidents by Hour of day

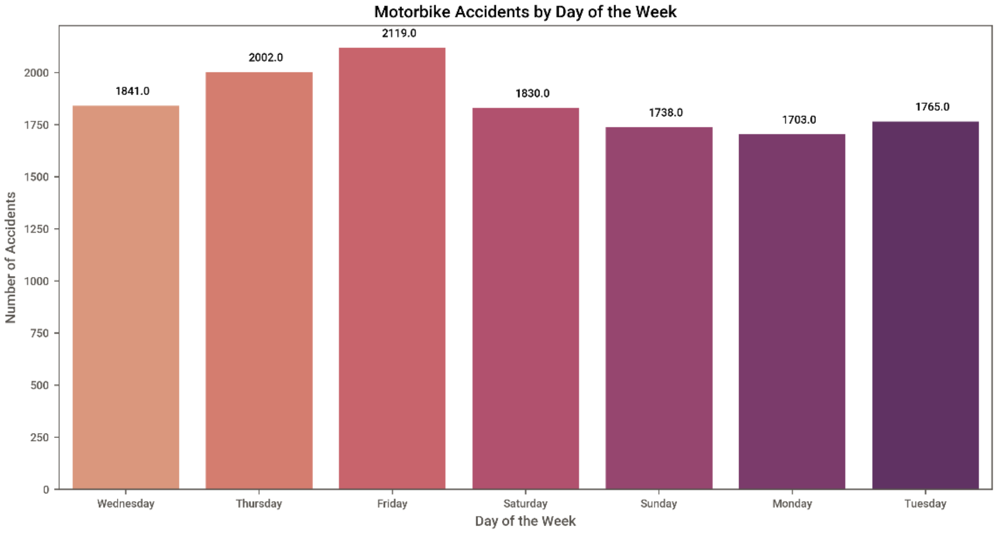

Figure 5: Motorbike Accidents by Day of Week

Analysis revealed that accidents involving pedestrians occur more in the afternoon around 15:00 and more frequently during the week compared to weekends, as shown in Figures 6 and 7. According to the United Kingdom Department for Transport (2023) factsheet, 14,750 pedestrian accidents occurred in 2020, which is the lowest in 16 years.

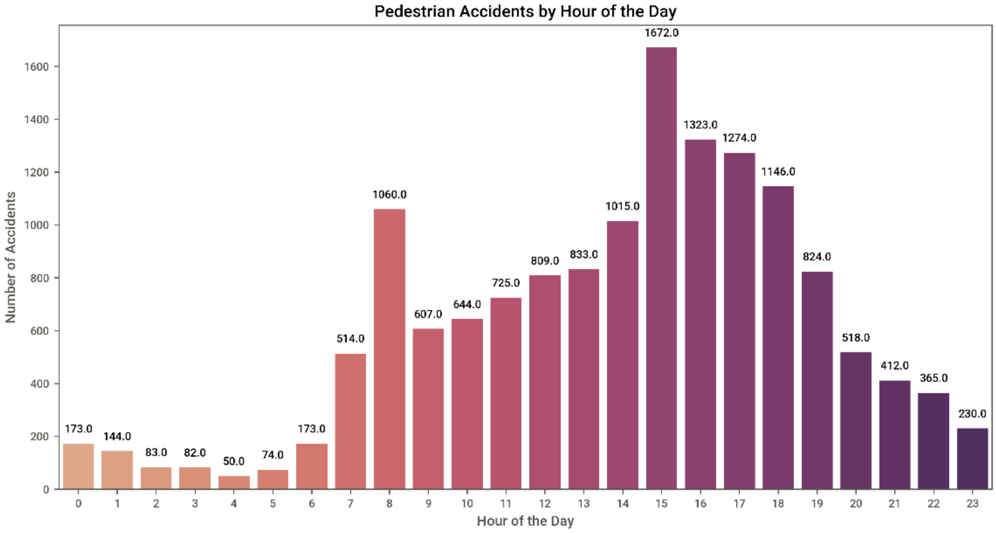

Figure 6: Pedestrian Accidents by Hour of Day

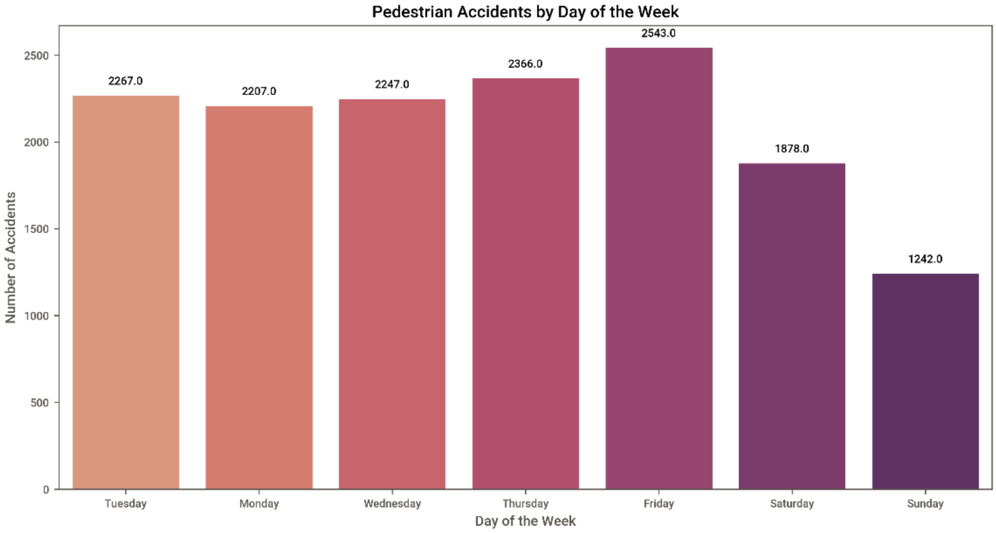

Figure 7: Pedestrian Accidents by Day of Week

Using the Apriori algorithm, explore the impact of selected variables on accident severity.

According to Zhang and Zhang (2002), association rule mining involves finding a pattern or relationship among factors (itemsets). In an application, it enables us to discover items that co-occur by specifying rules. Srikant (1996) posited that fast algorithms exist that can significantly improve mining associations between itemsets in large datasets using the Apriori algorithm. Thus, the Apriori algorithm was utilized to examine the influence of filtered variables on accident severity. The minimum threshold for support was increased due to computational issues being a drawback of the Apriori algorithm, as supported by Srikant (1996). Table 1 shows the threshold values for various rules.

Table 1: Association Mining Rules

Rule	Threshold (%)
Support	0.5
Confidence	0.7
Lift	1.25
Conviction	1.5

The justification for choosing 50% support is that itemsets occur at least half the time in the database. Furthermore, with a confidence of 70%, it means the rule is reliable. In other words, when the antecedent conditions of the rules are met, the consequent occurs 70% of the time. Finally, the lift reveals that 25% of occurrences are not due to chance, revealing a meaningful independent relationship. In other words, the consequent occurs more than 25% of the time when the antecedent is present.

The **association rules** reveal that **drivers** are the most frequent casualties in accidents, particularly those classified as **slight accidents**. Importantly, these accidents **do not involve pedestrians** directly. In essence, the rules suggest that **driver actions**, rather than pedestrian interactions or infrastructure (like crossings or pedestrian movement), play a significant role in the occurrence of these **slight accidents.**

2.3 Spatial Analysis: Where Do Accidents Happen?

Three regions were filtered using structured query language commands, particularly the East Riding of Yorkshire, Humberside, and Kingston Upon Hull. Geographic clustering using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied due to its ability to detect clusters of arbitrary shapes and effectively manage outliers. Also, it is a commonly applied spatial clustering algorithm (Pavlis et al., 2018).

In Figure 8, the three broad regions are separated into clusters using the longitude and latitude, which simply reflects the clustering of accidents within these regions.

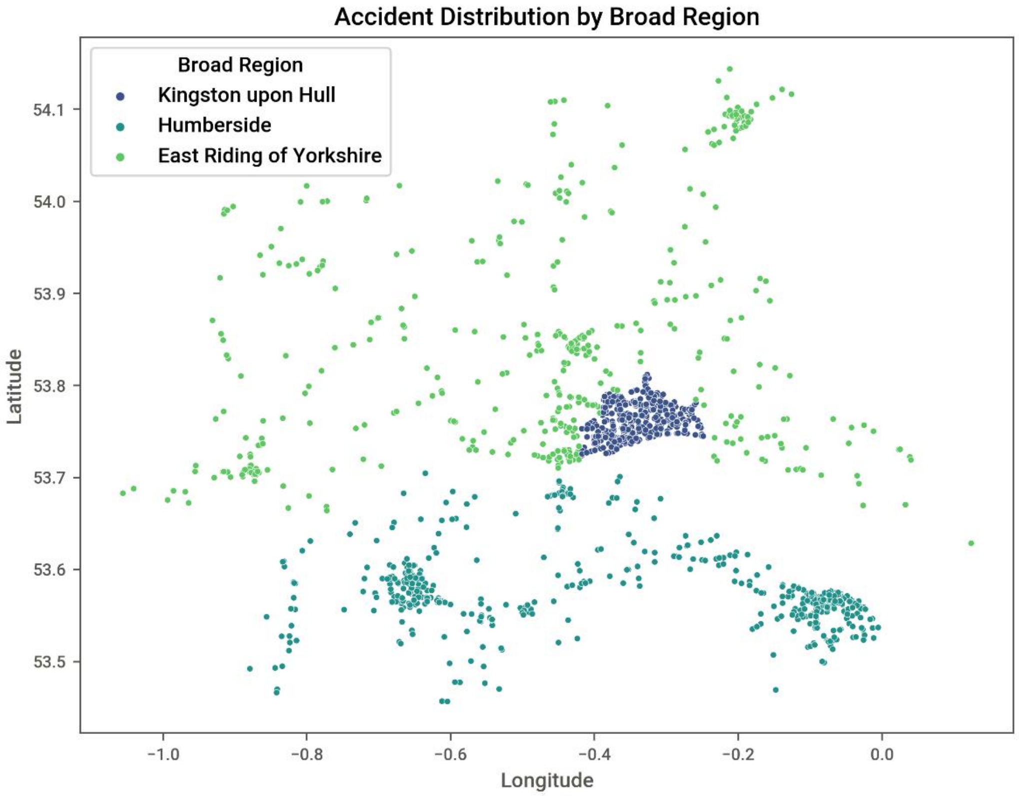

Figure 8: Cluster of Accident Distribution by Region

2.4 Under What Conditions do Accidents Happen?

Furthermore, discovering patterns and meaning from unlabelled data is an integral part of data mining employing cluster analysis (Ikotun et al., 2023). The K-means algorithm was chosen for its simplicity and low computational complexity. It is useful for finding relationships between datapoints. Principal component analysis was applied to reduce the dimensionality of the data, enabling working with the most important features as corroborated by Greenacre et al. (2022).

As shown in Figure 9, cluster analysis reveals accidents frequently occur during fine, rainy, and combined fine and windy weather conditions with varying speed limits from 30km/hr upwards.

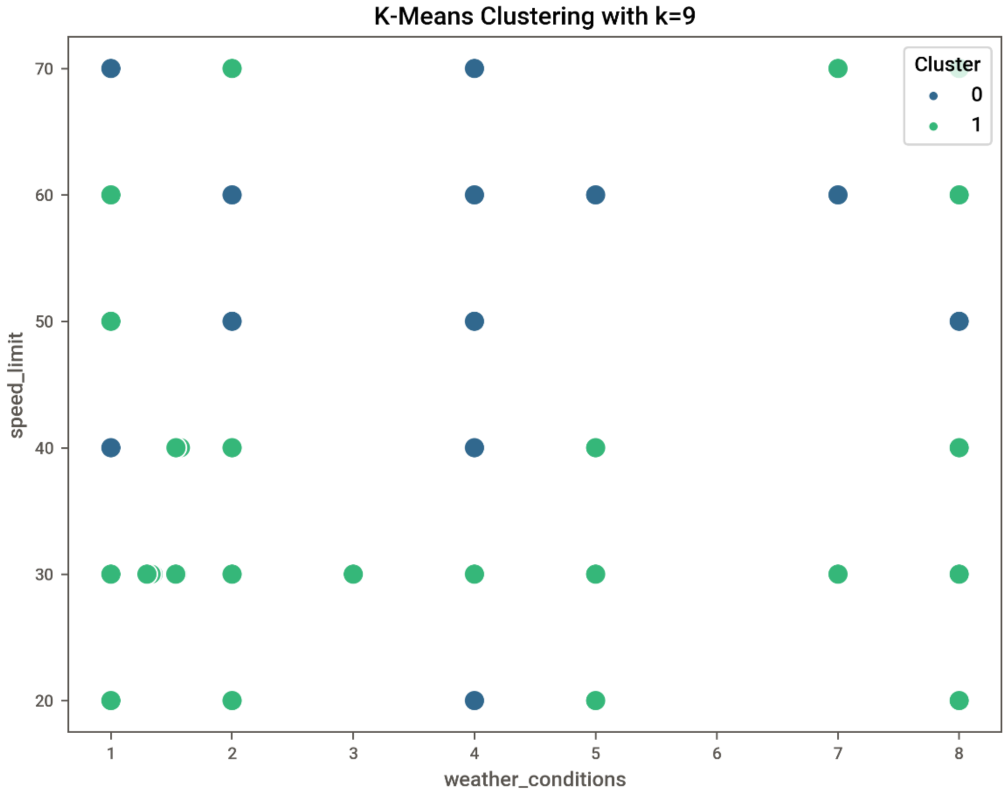

Figure 9: Speed Limit and Weather Condition clusters

Similarly, vehicle type and accident severity were clustered. As shown in Figure 10, vehicle types encoded between 1-23 (See road safety open dataset guide for reference) were involved in more accidents than other vehicle types, with the prevalent accident severity being slight.

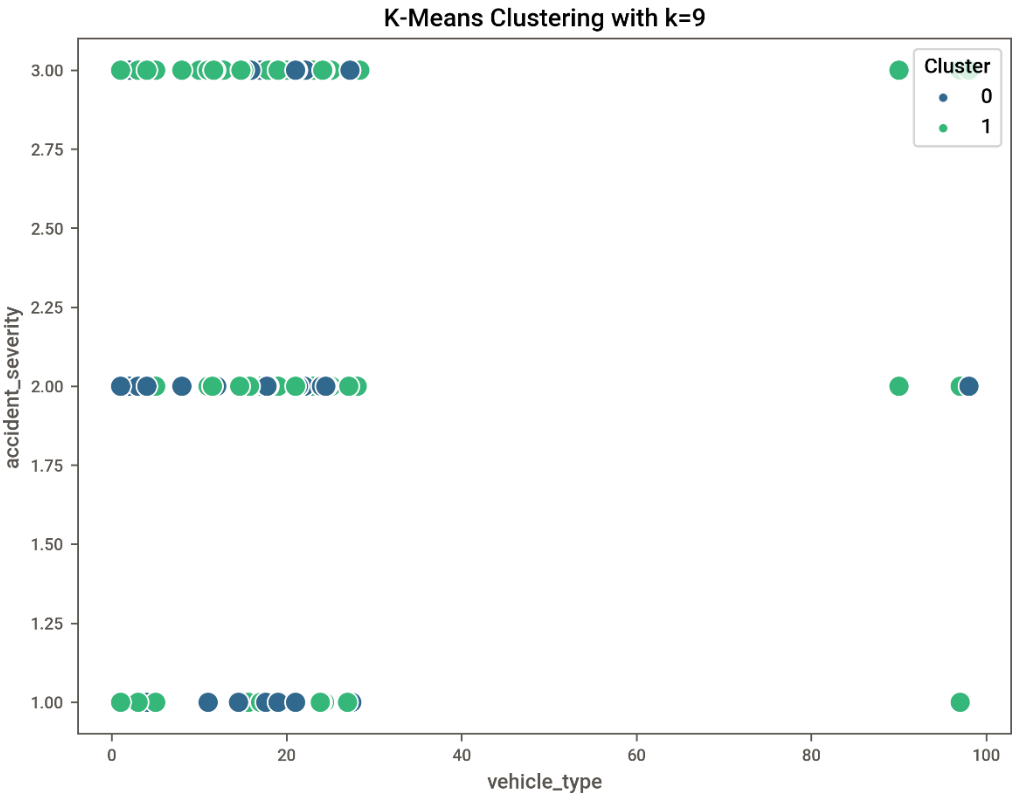

Figure 10: Speed Limit and Weather Condition clusters

2.5 Social Network Analysis

Social network analysis offers a broad approach to investigating the strength of relationships between individuals in a network (McLevey et al., 2024). This informs our understanding of social patterns of behaviours and thoughts. In this social network analysis, Facebook data has been analysed to identify the total number of individuals or entities in the network, the number of relationships between the nodes, the strength of connections in the network, and the average number of connections per node. The network is shown in Figure 11.

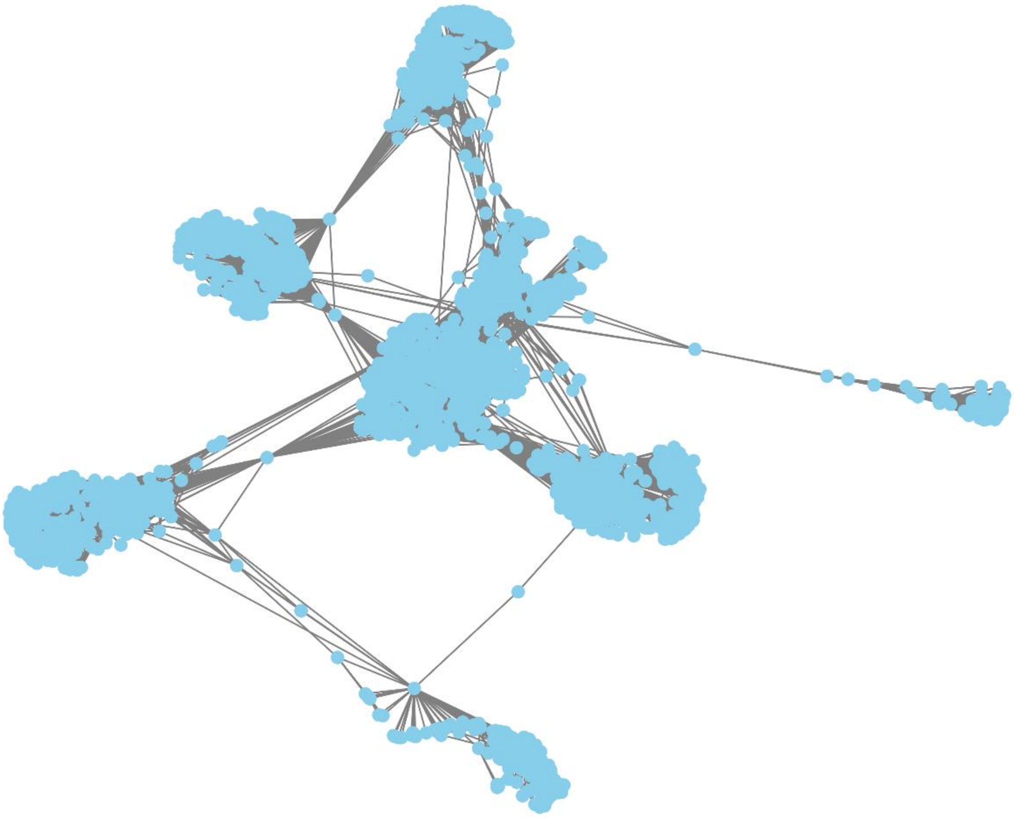

Figure 11: Social Network Visualization

Analysis revealed that there are 4039 nodes in the network and 88234 edges between them. The network density was 0.0108, which indicates a sparse relationship. However, the average degree of the network is 43.69, implying that each node in the network is connected to many other nodes, which can facilitate information spread and offer an influential communication pipeline.

To identify the most influential nodes within a graph network, we utilize centrality measures (Laghridat & Essalih, 2023). As such, the edge centrality of the network was also investigated. The betweenness and closeness centrality of the network were evaluated. The betweenness centrality assists us in identifying the frequency of a node when positioned on the shortest path connecting other nodes, while closeness centrality indicates the proximity of a node to others in the network using the shortest path. As a result, Node 107 had the highest betweenness and closeness value in the network. This indicates that node 107 is the closest to all other nodes in the network and acts as a bridge between different parts of the network, as shown in Figures 12 and 13. Therefore, Node 107 is influential in communicating critical road accident safety measures across the network.

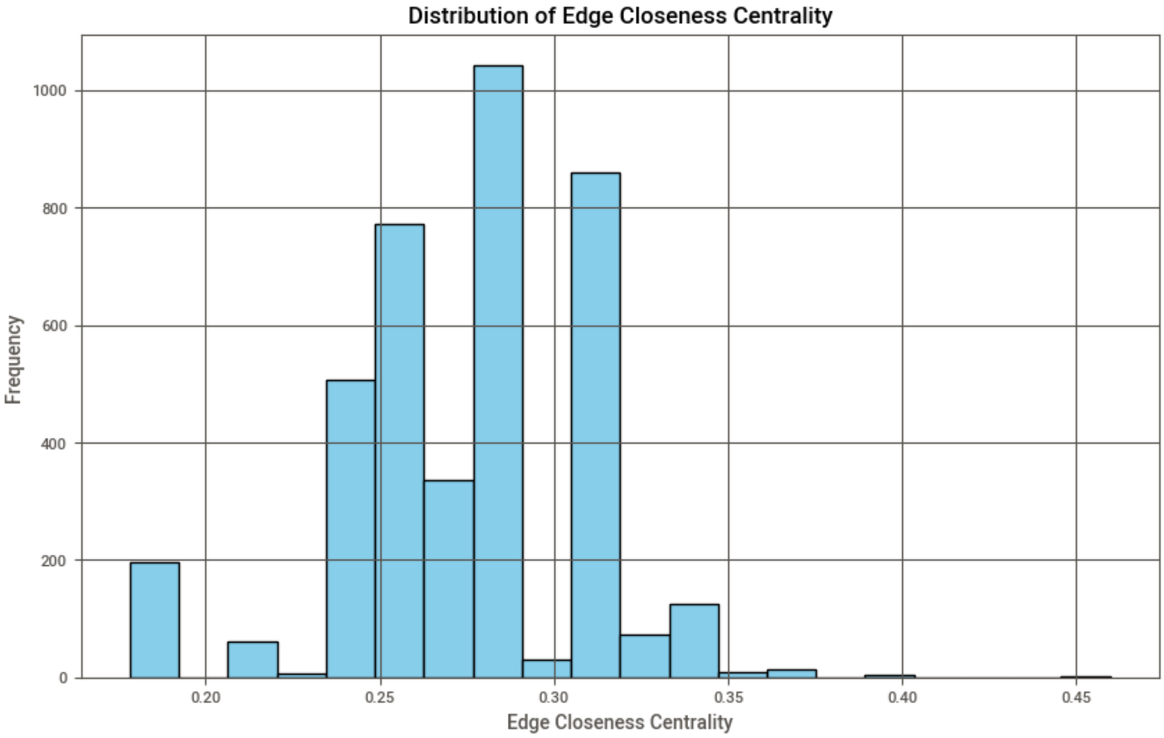

Figure 12: Closeness Centrality

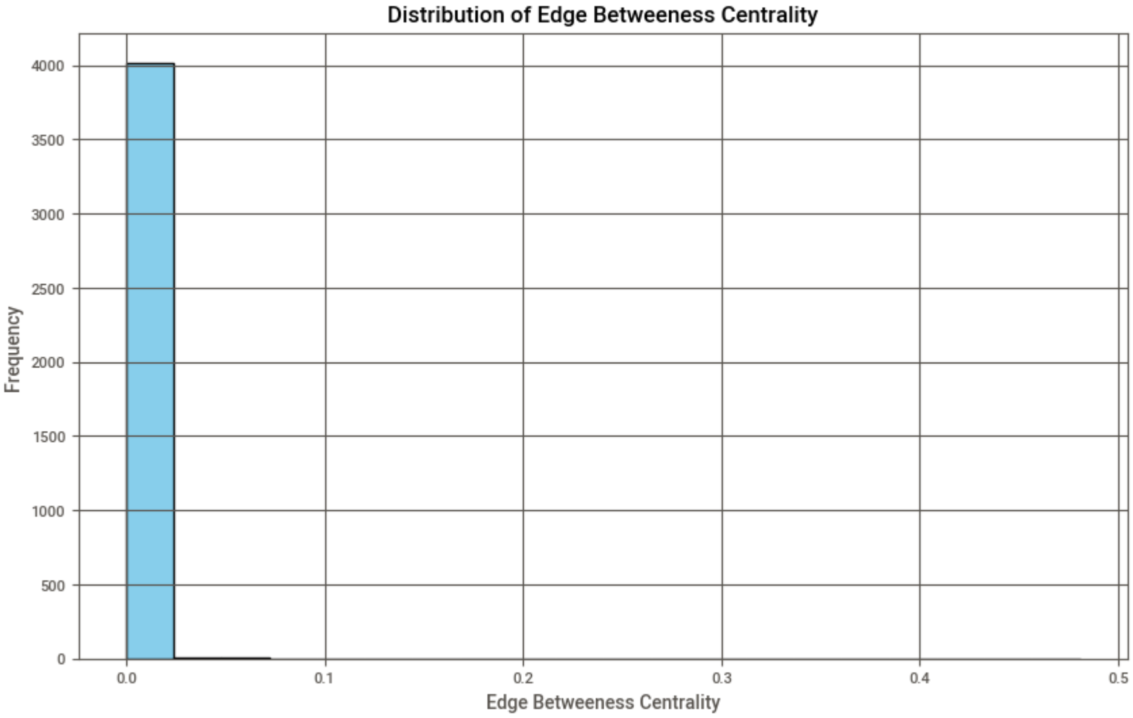

Figure 13: Betweenness Centrality

Two community clustering algorithms were applied to detect clusters within this social network. Louvain Modularity was chosen for its ability to detect hierarchical structures in communities by grouping smaller communities into larger ones (Blondel et al., 2024). The label propagation algorithm requires no a priori conditions and is suitable for community detection in large-scale networks. However, one of its drawbacks is that each node is considered equally significant, disregarding assigned weight (Liu et al., 2024).

Table 2: Comparison of Community Clusters Algorithms

Metrics	**Louvain Modularity Algorithm**	**Label Propagation Algorithm**
Number of Communities	13	44
Average Community Size	310.7	91.8
Modularity	0.78	0.73

As shown in Table 2, Louvain modularity formed 13 community clusters, with a high modularity score of 0.78. Although the largest cluster has 983 nodes and the smallest has 6 nodes, this reveals an imbalance and may be concluded that Louvain tends to prioritize strong modular structures while ignoring smaller noise-like groups.

In comparison, Label propagation formed 44 community clusters, with a lower modularity score of 0.73; however, still high. While one community has 1030 nodes, most of the others are small, which points to the fact that the algorithm detects many small communities, potentially being more sensitive to local patterns and over-segmentation.

Figure 14 shows the visualization of the community structures. Revealing similarities and differences. From the image, we can see how well the network is partitioned into communities of similar interests.

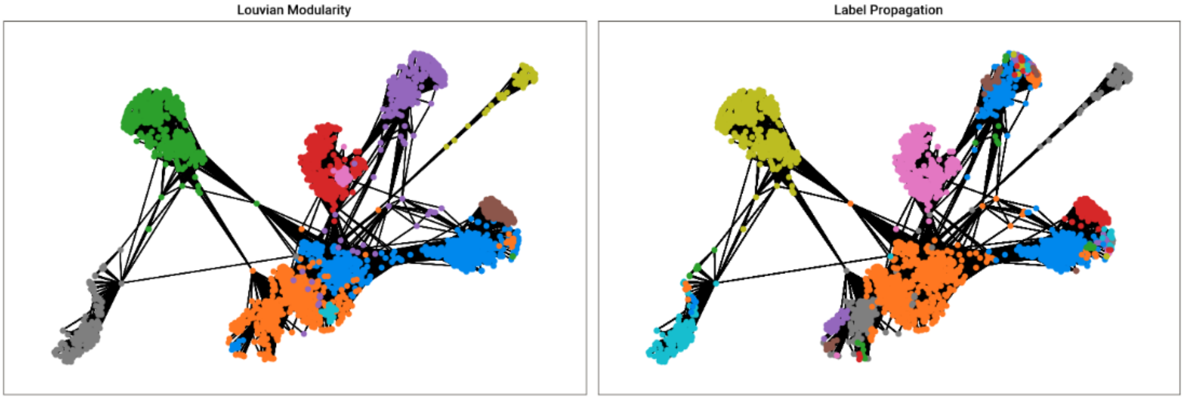

Figure 14: Community Clusters Visualization

PREDICTIONS

To predict weekly accidents for selected police force regions and daily accidents in high-risk Lower Layer Super Output areas (LSOAs) in Kingston upon Hull, time series analysis was employed. A time series is a chain of data points recorded at successive points in time, typically at uniform intervals, i.e., daily, monthly, yearly. It aids the forecasting of future occurrences by examining past occurrences and their autocorrelation (Deistler and Scherrer, 2022).

For weekly accident predictions across the three chosen regions, namely, Surrey, Kent, and Essex, appropriate forecasting model selection was imperative. This was done by examining the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots to detect significant lag relationships. However, while ACF and PACF are critical to determine predictive modelling, they are difficult to interpret and may be sensitive to noise (Alzawbaee & Mahmmood, 2024). Hence, an alternative model parameter selection method was chosen, the Akaike Information Criterion (AIC).

The data showed the presence of seasonality, indicating that a seasonal autoregressive integrated moving average model (SARIMA) would be appropriate. However, to ensure robustness it was compared to an autoregressive moving average (ARMA), and a machine-learning based XGBoost model. XGBoost, an efficient machine learning algorithm based on gradient boosting that improves predictive accuracy by combining multiple decision trees with parallelization, and optimization techniques (Chen & Guestrin, 2016). While its potential for overfitting with noisy data can be a drawback, its ability to capture complex relationships and deliver high predictive accuracy is a key strength.

Model hyperparameters were selected using the Akaike Information Criterion (AIC). It is essentially an estimator of prediction error given a set of statistically modelled data (Sutherland et al., 2023), and aids in selecting the best-fitting model.

Furthermore, the Augmented Dickey-Fuller test (ADF) was conducted to check for stationarity, which revealed that differencing was not required for either of the models. The ADF test helps examine the presence of a unit root (Sultan, 2023). Stationarity is important to stabilise the mean and variance of a time series model.

The performance of each model was evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), as summarized in **Table 3**.

Table 3: Weekly Prediction Time Series Model Performance Summary

LSOA	Model	Forecast Granularity	RMSE	MAE	MAPE (%)
Surrey	**SARIMA**	**weekly**	**17.18**	**13.70**	**42**
	ARIMA	**weekly**	**19.78**	**16.45**	**50**
	XGBOOST	**weekly**	**16.08**	**13.04**	**36.93**
Kent	**SARIMA**	**weekly**	**26.65**	**20.52**	**51**
	ARIMA	**weekly**	**28.28**	**21.90**	**54**
	XGBOOST	**weekly**	**23.61**	**19.05**	**40.52**

Essex	**SARIMA**	**weekly**	**26.27**	**21.19**	**53**
	ARIMA	**weekly**	**23.59**	**18.78**	**47**
	XGBOOST	**weekly**	**22.45**	**18.62**	**41.10**

As shown in Table 3, XGBoost outperformed the statistical models across the three regions in predicting weekly accidents for the upcoming year, 2020. XGBoost exhibited significantly lower RMSE, MAE, and MAPE values, indicating better overall predictive accuracy. The XGBoost predictions are shown in Figures 15, 16, and 17.

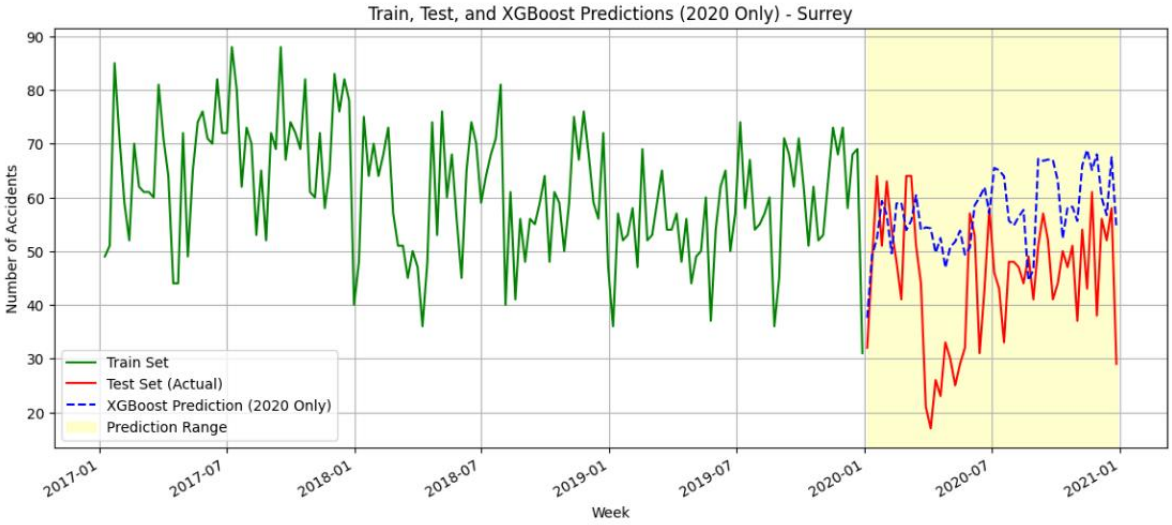

Figure 15: XGBoost Prediction for Surrey Region

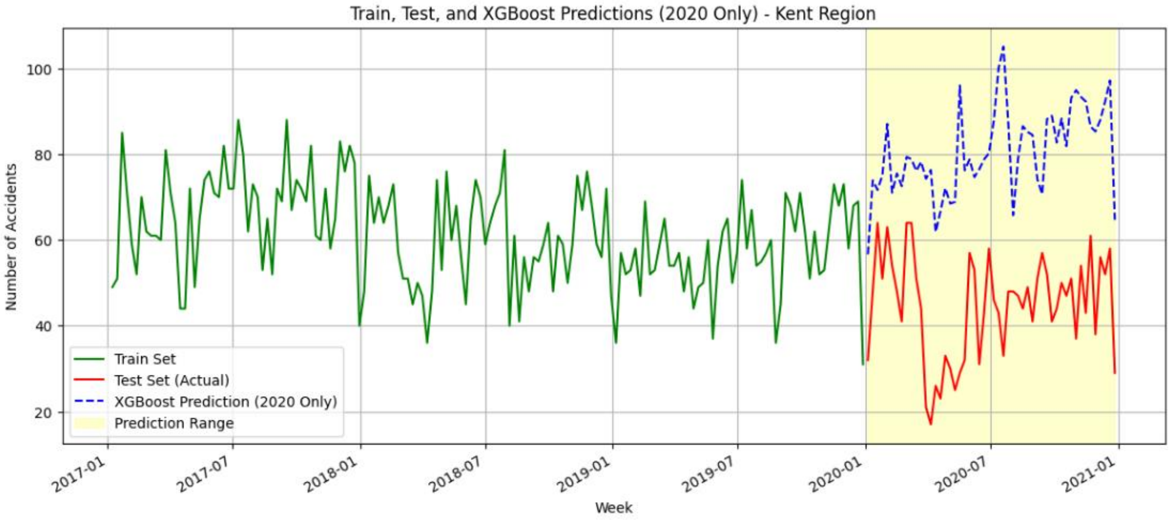

Figure 16: XGBoost Prediction for Kent Region

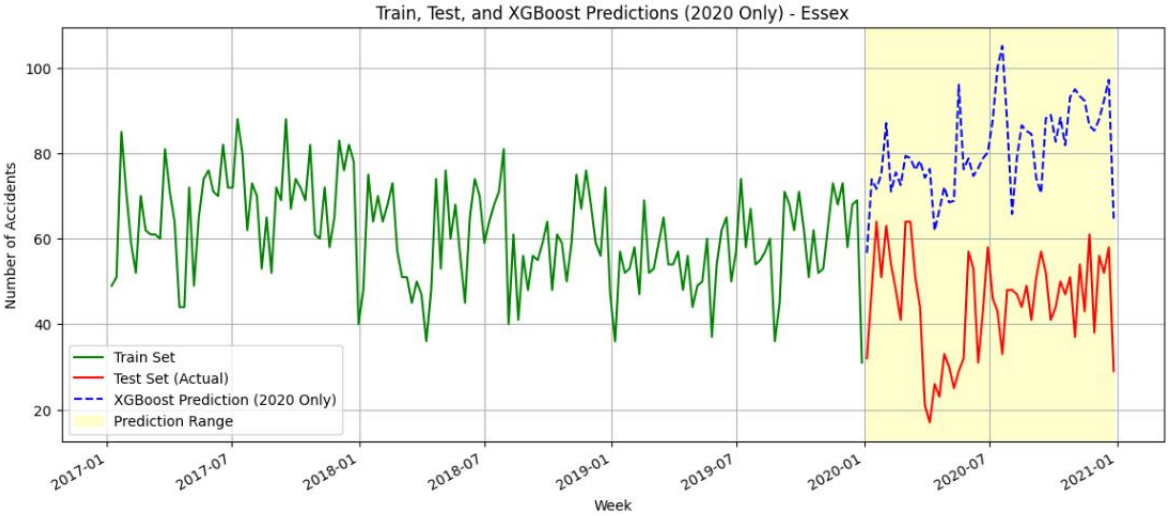

Figure 17: XGBoost Prediction for Essex Region

For daily accident prediction, the three lower layer super output areas in Kingston Upon Hull with the highest accident count are E01012817, E01012889, and E01012848, with 10, 7, and 7 accidents count respectively from January to March in 2020. The SARIMA model was chosen to forecast the daily accident prediction across the three areas for July. The model forecasted mostly zero accidents across all three areas for July.

RECOMMENDATION

Based on the extensive analysis conducted, recommendations are suggested to the Government to improve traffic control and reduce accident incidents. The recommendations are:

- **Implement Dynamic Traffic Control Systems:** It is suggested that during weekdays when it is busiest, dynamic traffic control measures such as AI-driven systems be introduced in high incident areas to mitigate erratic driving behaviour during peak hours.
- **Community-Based Safety Program:** Local communities could be mobilized towards safety through influential community members. Targeted safety campaigns for pedestrians to increase their awareness during peak hours on weekdays. Campaigns such as improved road signage, media publicity, and community outreach to reinforce safe road crossing practices.
- **Improved Vehicle License:** Driving evaluation programs could be revamped to improve the driving ability of drivers. Also, certain vehicle types should have additional testing programs. During peak hours on weekdays drivers should be taught to exercise more patience and adopt increased safe driving behaviours.
- **Data-Driven Infrastructure Planning:** High incident areas can be improved by prioritizing road maintenance and implementing safety cars during specific weather conditions to mitigate accident occurrences.

REFERENCES

Alzawbaee, M. and Mahmmood, O., 2024. Determine the Best Models for Time Series by using a New Suggested Technique. *General Letters in Mathematics (GLM)*, *14*(1).

Blondel, V., Guillaume, J.L. and Lambiotte, R., 2024. Fast unfolding of communities in large networks: 15 years later. *Journal of Statistical Mechanics: Theory and Experiment*, *2024*(10), p.10R001

Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Deistler, M. and Scherrer, W., 2022. *Time series models*. Cham: Springer.

Department for Transport (2023) *Reported road casualties in Great Britain: pedestrian factsheet, 2022*. Available at: Reported road casualties in Great Britain: pedestrian factsheet, 2022 - GOV.UK [Accessed: 7 April 2025].

Gov.uk (2022) *Reported road casualties in Great Britain: motorcyclist factsheet 2022*. Available at: https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-motorcyclist-factsheet-2022/reported-road-casualties-great-britain-motorcyclist-factsheet-2022 [Accessed 7 April 2025].

Greenacre, M., Groenen, P.J., Hastie, T., d'Enza, A.I., Markos, A. and Tuzhilina, E., 2022. Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), p.100.

Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J., 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, pp.178-210.

Laghridat, C. and Essalih, M., 2023. A set of measures of centrality by level for social network analysis. *Procedia computer science*, *219*, pp.751-758.

Liu, M., Yang, J., Guo, J. and Chen, J., 2024. A label propagation community discovery algorithm combining seed node influence and neighborhood similarity. *Knowledge and Information Systems*, *66*(4), pp.2625-2649.

McLevey, J., Scott, J., Carrington, P.J., Prell, C. and Schaefer, D.R., 2024. Introducing Social Network Analysis. In *The Sage Handbook of Social Network Analysis*. Sage Publications Ltd.

Pavlis, M., Dolega, L. and Singleton, A., 2018. A modified DBSCAN clustering method to estimate retail center extent. *Geographical Analysis*, *50*(2), pp.141-161.

Prakash, S., Singh, S. and Mankar, A., 2024, July. Bridging data gaps: A comparative study of different imputation methods for numeric datasets. In *2024 International Conference on Data Science and Network Security (ICDSNS)* (pp. 1-7). IEEE.

Srikant, R., 1996. *Fast algorithms for mining association rules and sequential patterns*. The University of Wisconsin-Madison.

Sultan, M.A., 2023. Forecasting the GDP in the United States by using ARIMA Model. *Can. J. Bus. Inf. Stud*, *5*(3), pp.63-69.

Sutherland, C., Hare, D., Johnson, P.J., Linden, D.W., Montgomery, R.A. and Droge, E., 2023. Practical advice on variable selection and reporting using Akaike information criterion. *Proceedings of the Royal Society B*, *290*(2007), p.20231261.

Zhang, C. and Zhang, S. eds., 2002. *Association rule mining: models and algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg.