Introduction

The aim of this report was to analyze the UKs Road accident data for the year 2020 to gain insights on how to advice government agencies on measures to take to improve road safety. The database used for the analysis consisted of four tables with records of road accidents and the main objective was to build a predictive classification model for accident severity.

Data Analysis

The database used for the analysis contained four tables which had several missing values and were addressed via data cleaning techniques. Additionally, external data sourced from (doogal.co.uk, n.d.) was used to fill used to fill in some missing values. A detailed record of the steps taken during the data cleaning phase can be reviewed in the attached Jupyter notebook.

Analysis of when accidents occur.

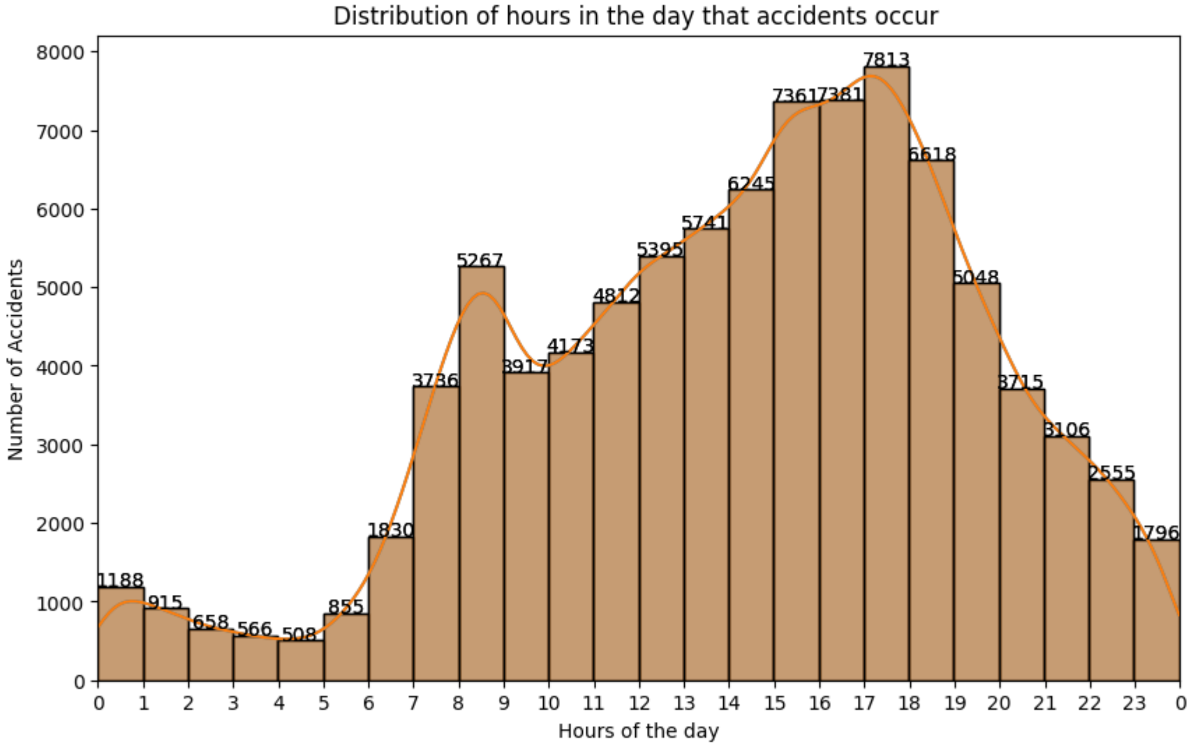


Figure 1 Distribution of hours in a day that accidents occur.

It can be seen in Figure 1 that there are two significant spikes in the distribution of hours of the day that accidents occur. The spikes show that accidents are likely to occur between 8:00 and 10:00 hours and are even more likely to occur between 16:00 to 18:00 hours. According to the article (More than half of car accidents happen during rush hour - Admiral.com, n.d.), road accidents in the UK tend to have a higher occurrence during morning and evening rush hours which are between 8:00

to 9:00 hours and 17:00 to 18:00 hours respectively, with the later having a higher number of accidents because drivers tend to be more stressed. This explains the two significant spikes seen in Figure 1 and why the spike in the between 16:00 to 18:00 hours have the highest accident occurrences.

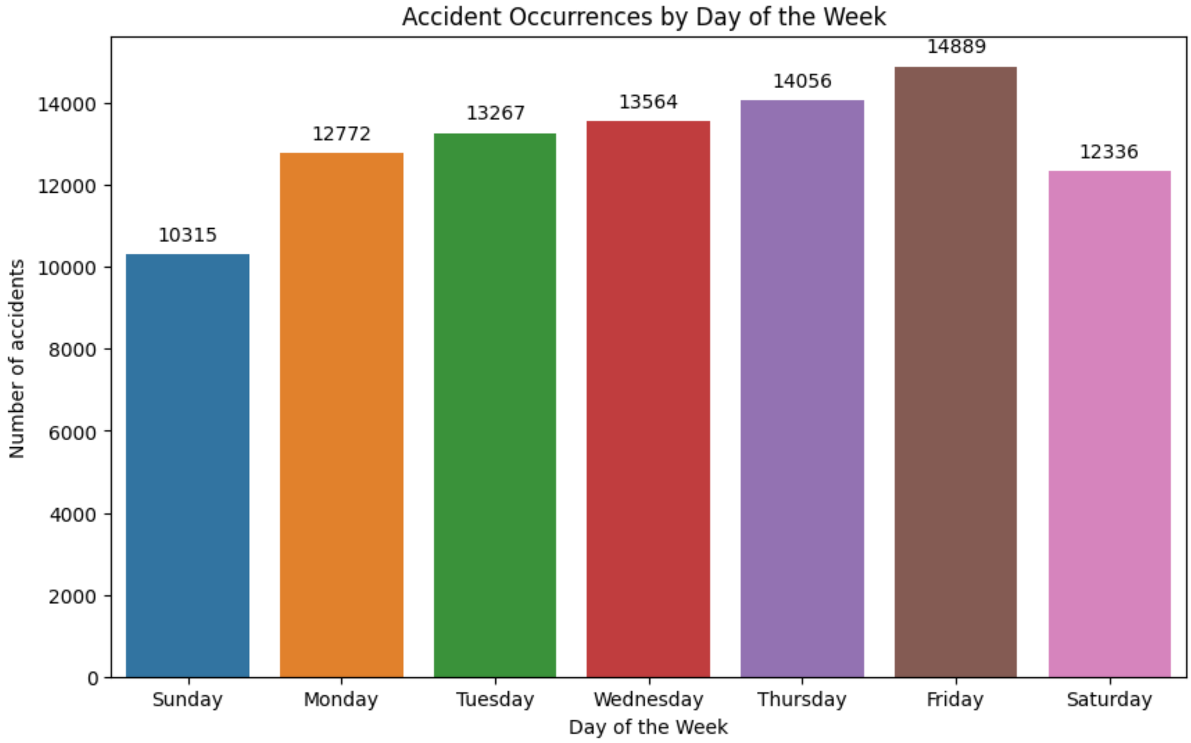


Figure 2. Count plot showing accident count for each day of the week.

It is seen in Figure 2 that the day with the most accidents in the week is Friday. According to the research (Epic, 2017) Thursdays and Fridays have the highest occurrence of accidents because towards the end of the week, drivers are often tired and, in a hurry, to get home and so they tend to drive riskier. This also explains why more accidents occur from 15:00 to 18:00 hours on Friday as seen in Figure 3.

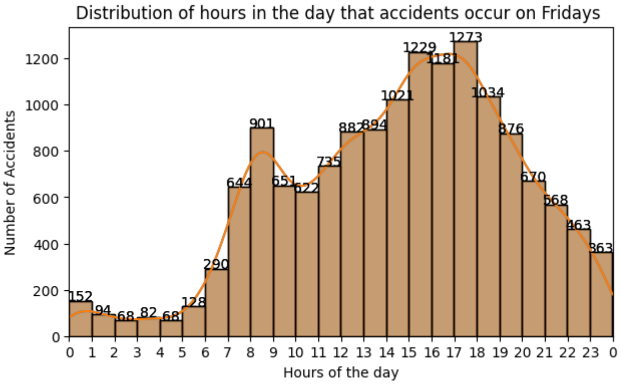


Figure 3. Distribution of hours that accidents occur in on Fridays.

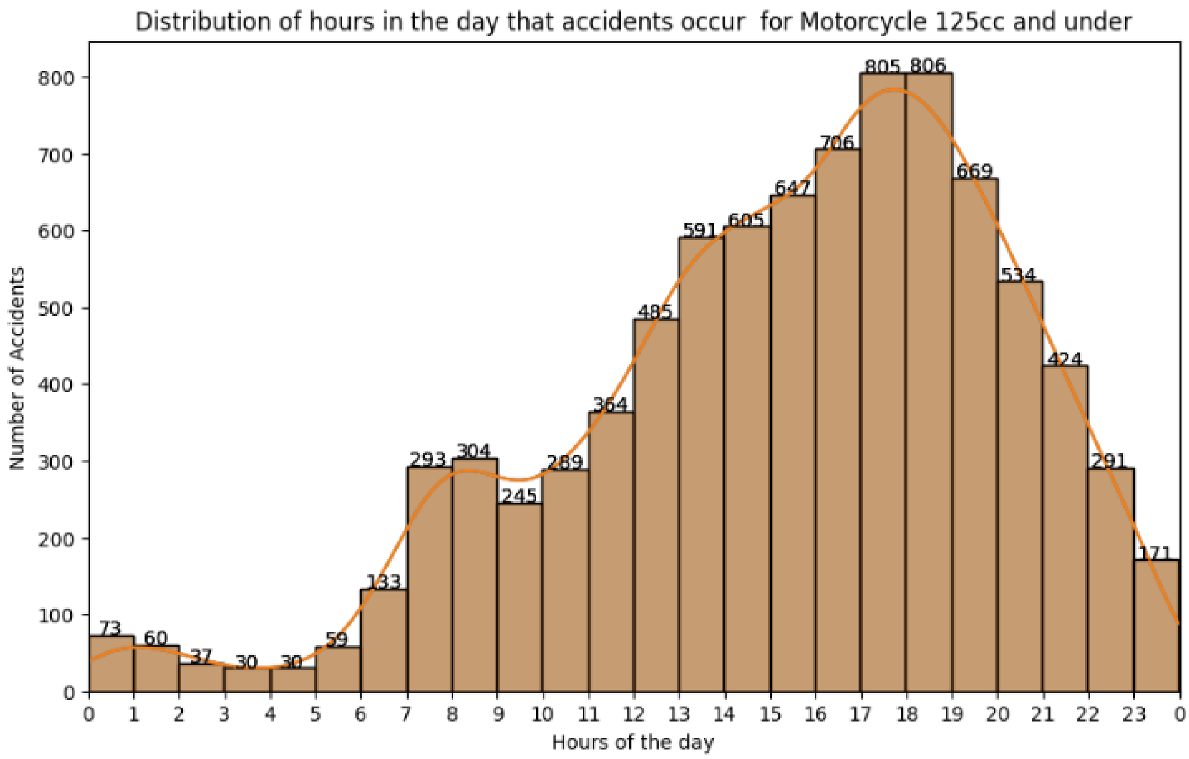


Figure 4. 2 Distribution of hours in the day that accidents occur for Motorcycle 125cc and under

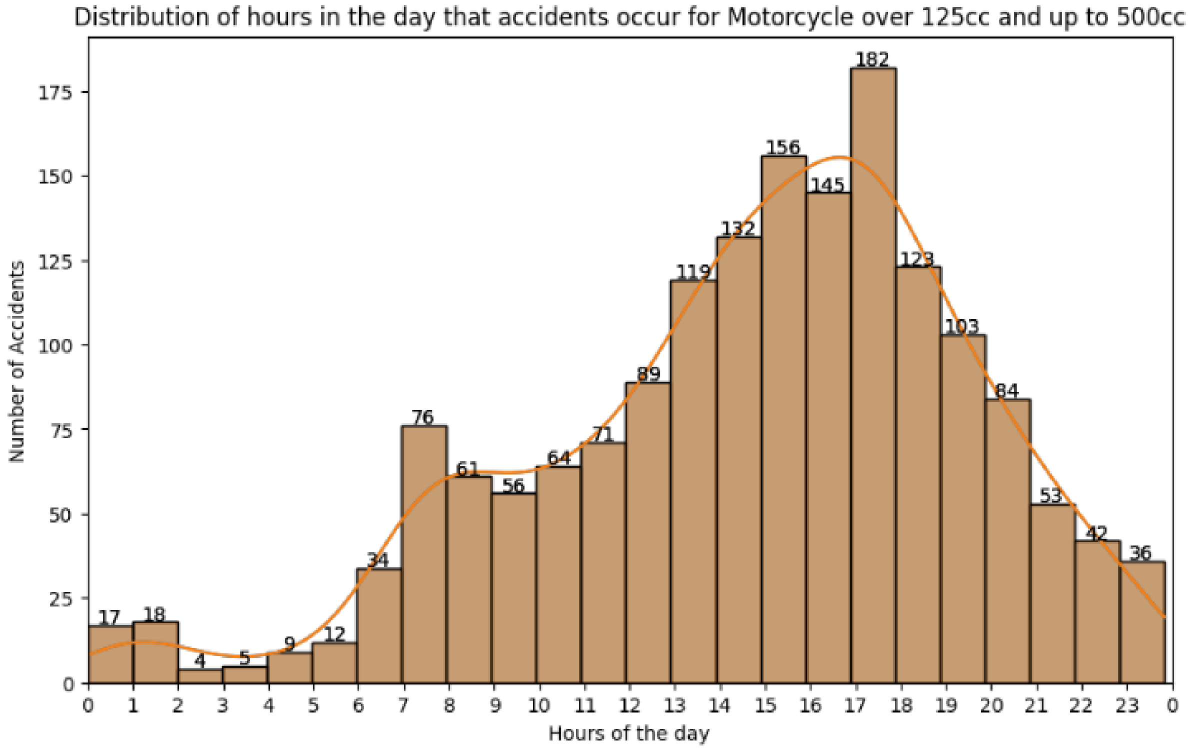


Figure 4. 1 Distribution of hours in the day that accidents occur for Motorcycle for 125cc and up to 500cc.

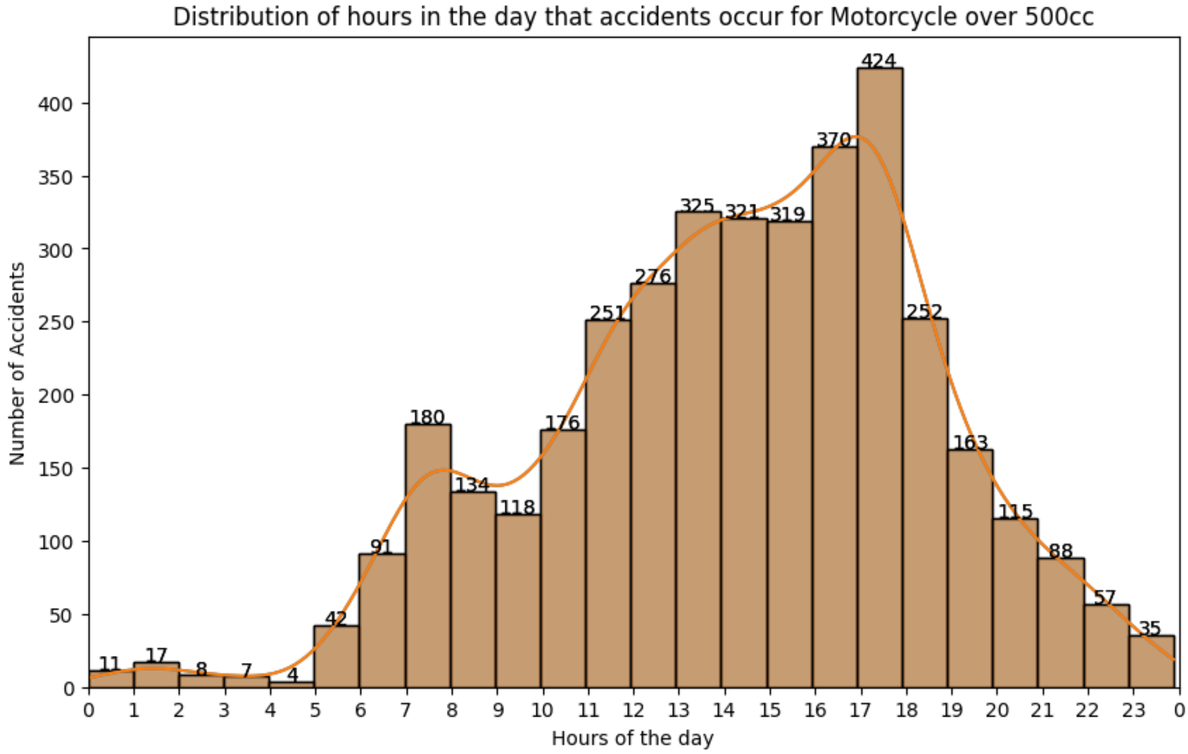


Figure 4. 3 Distribution of hours in the day that accidents occur for Motorcycle over 500cc.

For motorbikes, it can be seen from Figure 4. 1, Figure 4. 2 and Figure 4. 3 that accidents occur mainly between 17:00 to 18:00 hours. According to (Team L, 2021), accidents occur from 16:00 to 18:00 hours because of reasons like contending with the sunset and risky driving from drivers during rush hours trying to get home quickly.

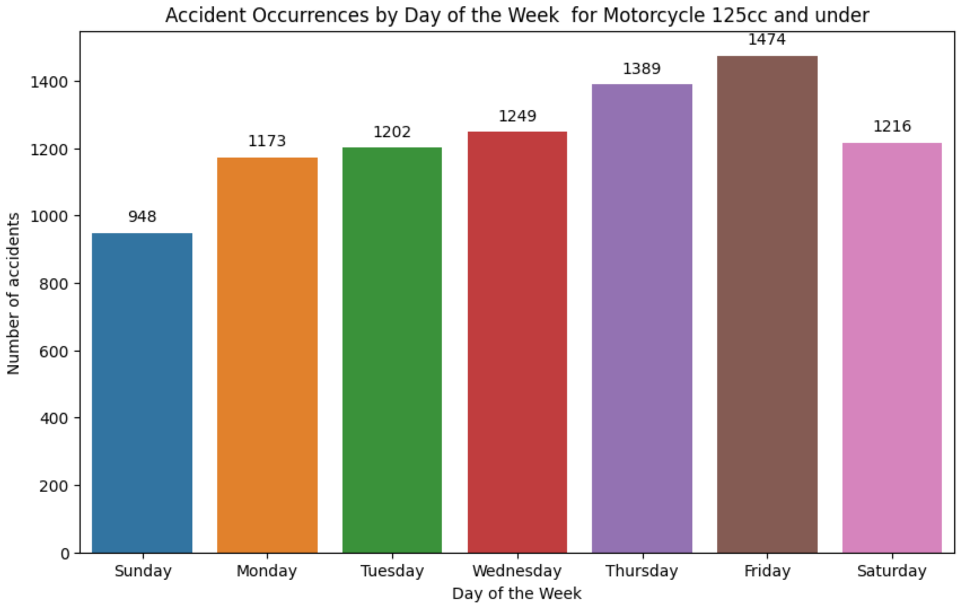


Figure 4 count plot showing accident occurrence for each day of the week for motorcycle 125cc and under

From Figure 4, Figure 5. 1 and Figure 5. 2, we deduce that motorbikes with engines below 500cc tend to have accidents on weekdays with Friday being the peak and for motorbikes over 500cc, accidents occur mainly on weekend with Sundays being the highest. According to the article (Swinton Insurance, 2020), motor bikes with an engine of over 500cc are used mostly for leisure and racing while the ones below 500cc are used for day-to-day jobs like food delivery, etc. and according to the book (Prideaux and Carson, 2010) people tend to ride motorcycles for leisure on weekends. This explains why accidents happen mainly on weekdays for motorbikes below 500cc and on weekends (especially Sunday) for motorbikes over 500cc.

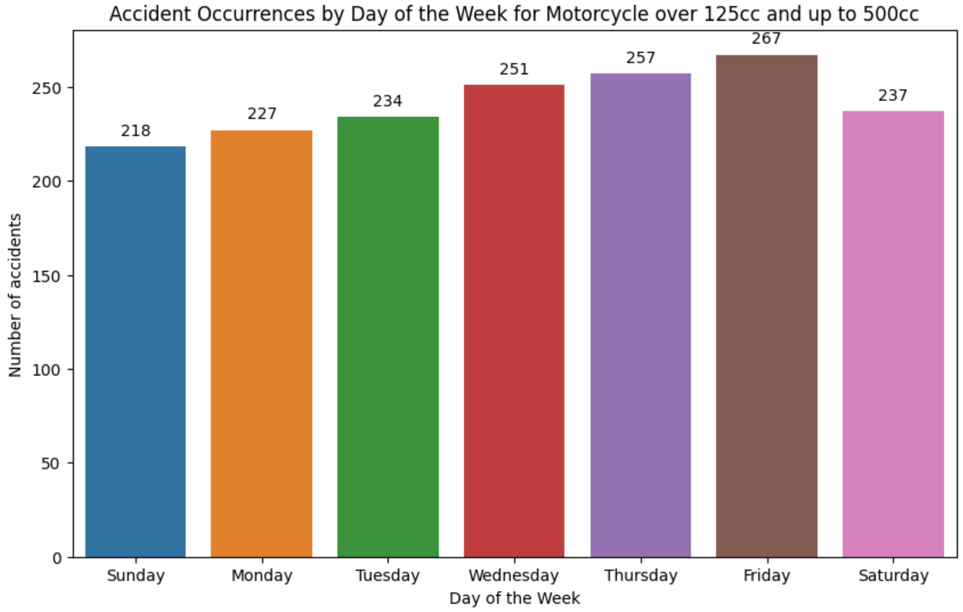


Figure 5. 1 count plot showing accident occurrence for each day of the week for motorcycle over 125cc and up to 500cc.

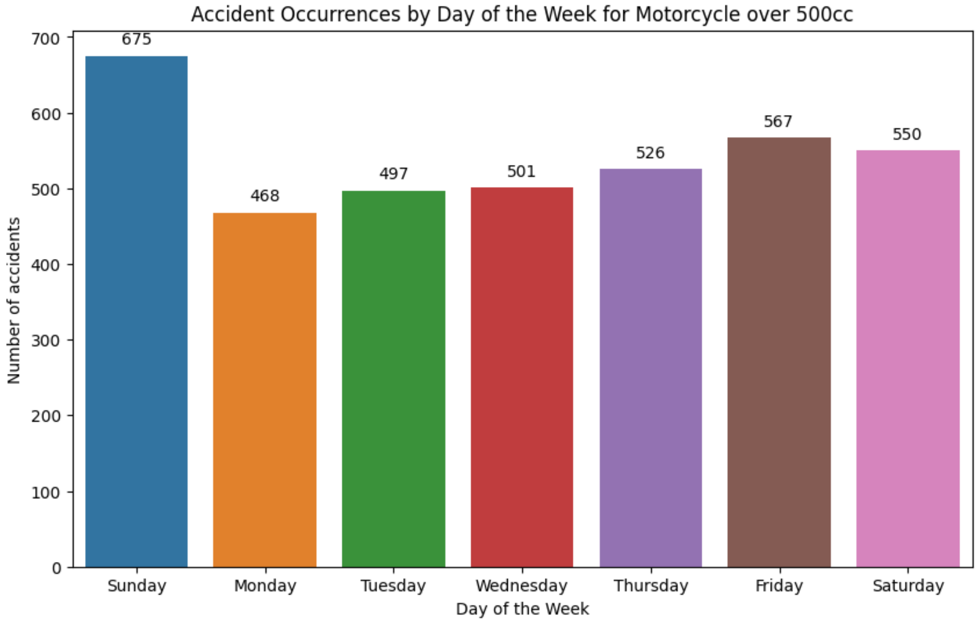


Figure 5. 2 count plot showing accident occurrence for each day of the week for motorcycle over 500cc.

It is seen from Figure 6 that accidents occurred the highest number of times on weekdays with Fridays being the highest and the least number of times on weekends with Sundays being the lowest for pedestrians. From comparing Figure 7 and Figure 8, we can agree with the article (Injury Facts, 2020) in saying that the reason why this occurs is because the roads are less busy during on weekends.

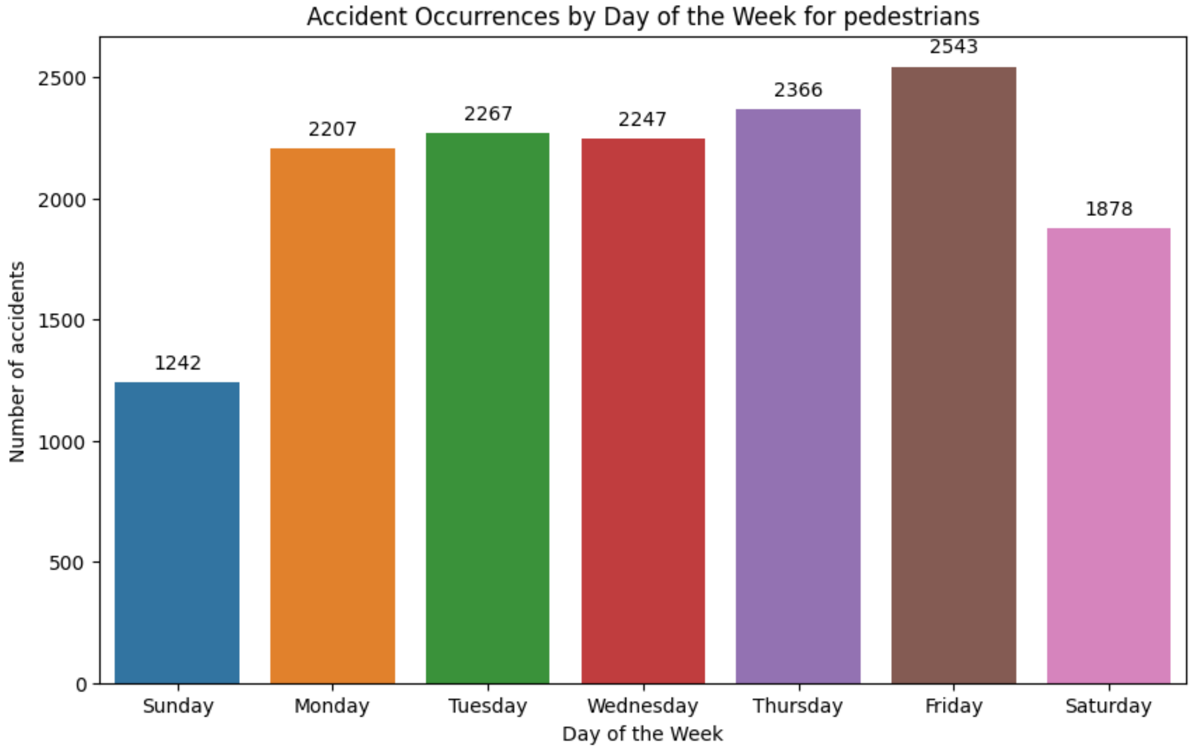


Figure 6 Count plot showing the number of accidents that occurred each day in 2020 involving pedestrians.

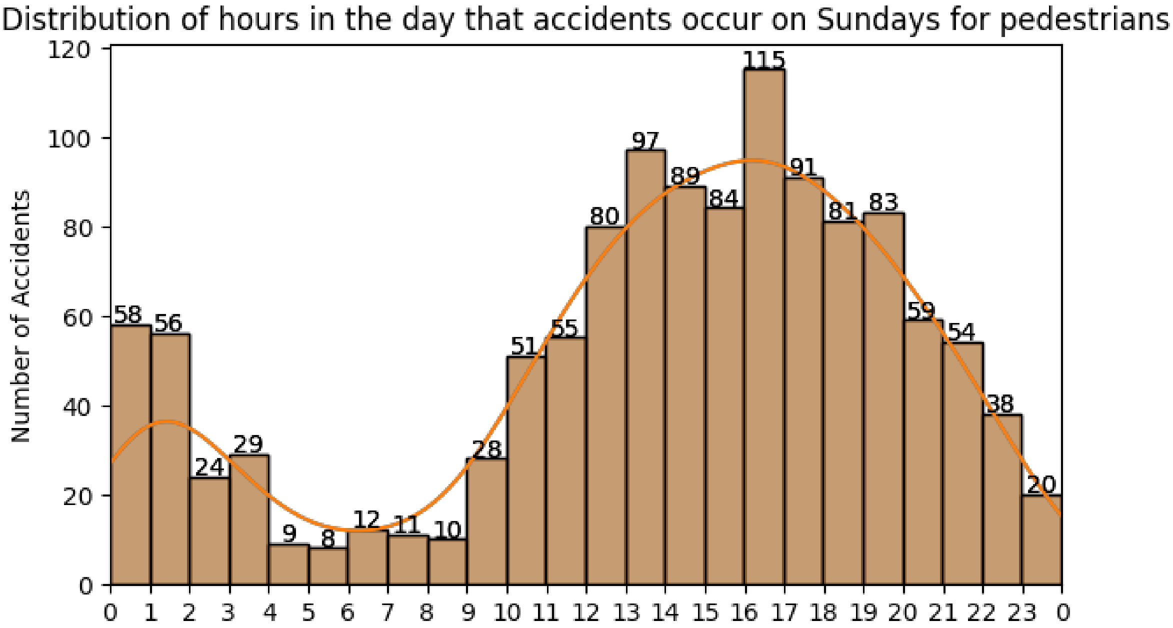


Figure 7 Distribution of hours in the day on Sundays that accidents involving pedestrians occurred.

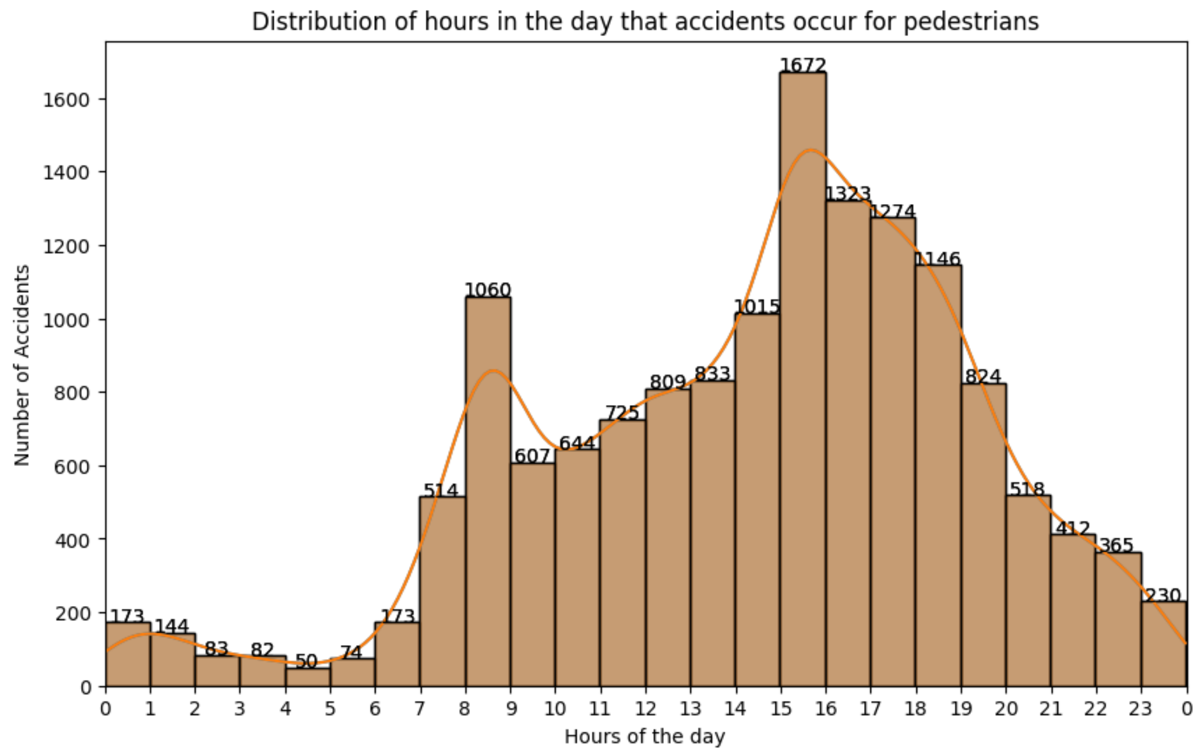
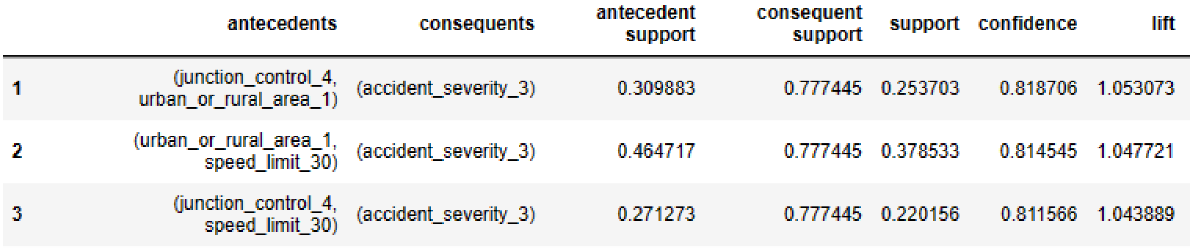


Figure 8 Distribution of hours in the day that accidents involving pedestrians occurred.

Analysis of the conditions that lead to accidents.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
1	(junction_control_4, urban_or_rural_area_1)	(accident_severity_3)	0.309883	0.777445	0.253703	0.818706	1.053073
2	(urban_or_rural_area_1, speed_limit_30)	(accident_severity_3)	0.464717	0.777445	0.378533	0.814545	1.047721
3	(junction_control_4, speed_limit_30)	(accident_severity_3)	0.271273	0.777445	0.220156	0.811566	1.043889

Table 1. Table showing selected rules generated by the apriori algorithm to showing the impact of selected variables on accident severity.

To answer the question “Under what conditions do accidents occur “, the apriori algorithm was used to generate rules to show the impact of selected variables on the severity of accidents. The most interesting rules have been selected and put in Table 1.

The first rule in Table 1 tells us that we are 82% confident that non-fatal accidents occur when there is a give way or no junction control in an urban area and this happened 25% of the time in the dataset.

The second rule in Table 1 tells us that when the speed limit in an urban area is 30mph, we are 81% confident that the accidents that occur are non-fatal and this happened 37% of the time in the dataset, and with a lift greater than 1, we know that the rule is useful. This rule is also familiar because according to research (Hunter et al., 2023), reducing accidents in urban areas has become a trend across Europe because studies have found that it reduces the rate of fatal injuries for both pedestrians and drivers.

The third in Table 1 rule was found to be interesting because it says that we are 81% confident that if there is no junction control but the speed limit is 30, then the accidents that occur are non-fatal and this occurred in 22 percent of the data, and we know that the rule is useful because the lift is greater than 1. This rule was found to be interesting because the research (Owen, 2019) argued, that fatal accidents often occur at busy roads with little or no junction control, which contradicts the rule. This gives the insight that speed limit is a is a very vital variable when considering accident severity.

Analysis of where accidents occur.

The analysis to determine where accidents occurred in the UK was carried out specifically for the Humberside region.

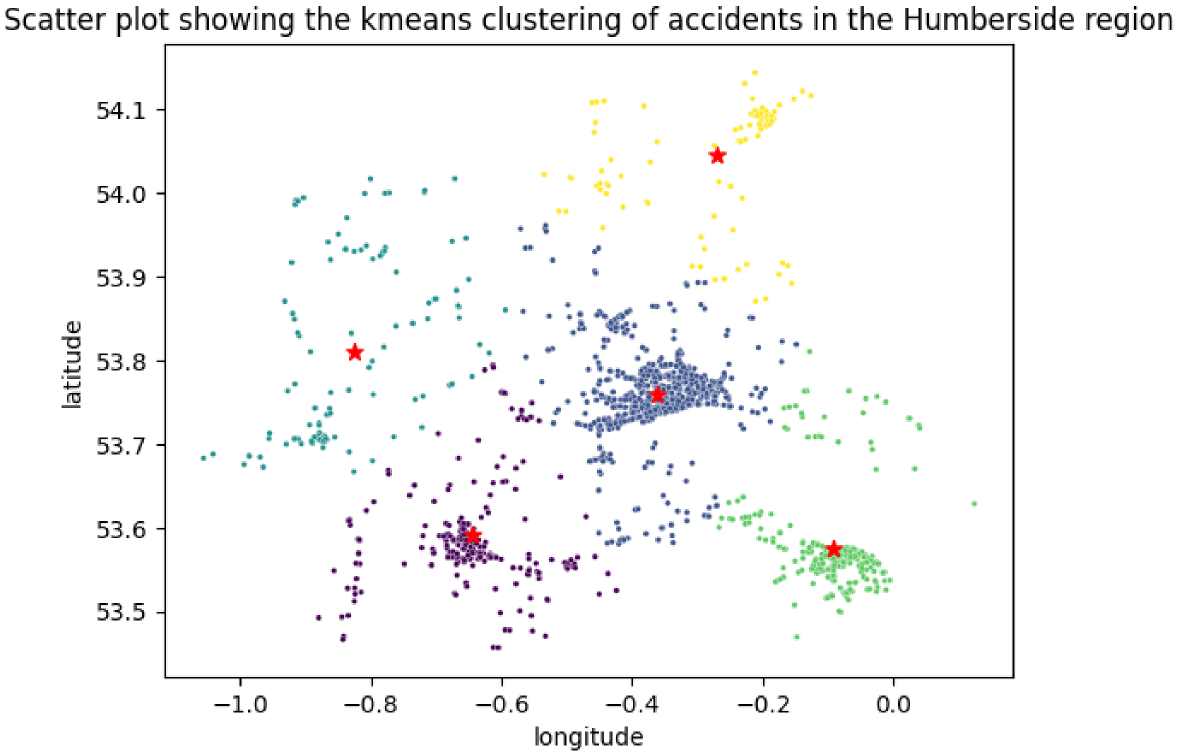


Figure 9 Scatter plot showing the clustering of accidents in the Humberside region.

It can be seen in Figure 9 that the densest cluster is the blue one in the middle and by comparing **Error! Reference source not found.** with it was evident that the area with the most accidents in Humberside in 2020 was hull.

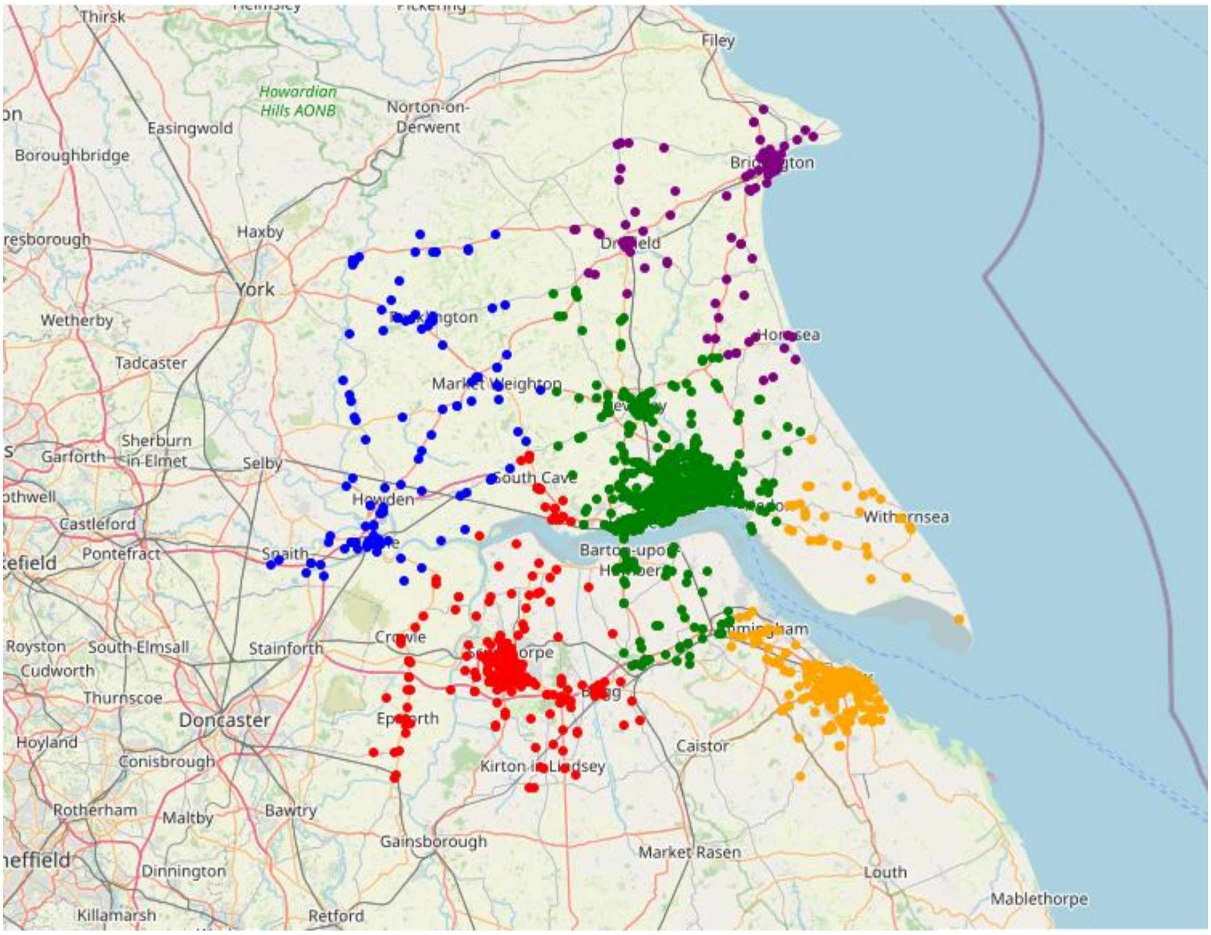


Figure 10 folium map showing the clustering of accidents in the Humberside region.

From Figure 11, Figure 12, Figure 13 and Figure 14 it can be deduced that accidents tend to occur more where there are major roads and according to the research (Young, 2018) this is because major roads are busier and as such will have more stressed drivers on the road leading to more accidents.

According to the article (Ford, 2022) the reason that hull has the highest number of accidents is not solely because it is in the center of the Humberside region, in fact it is stated in the article that hull has the most dangerous roads in the whole of the country hence leading to a higher accident rate than the other cities in the Humberside region.

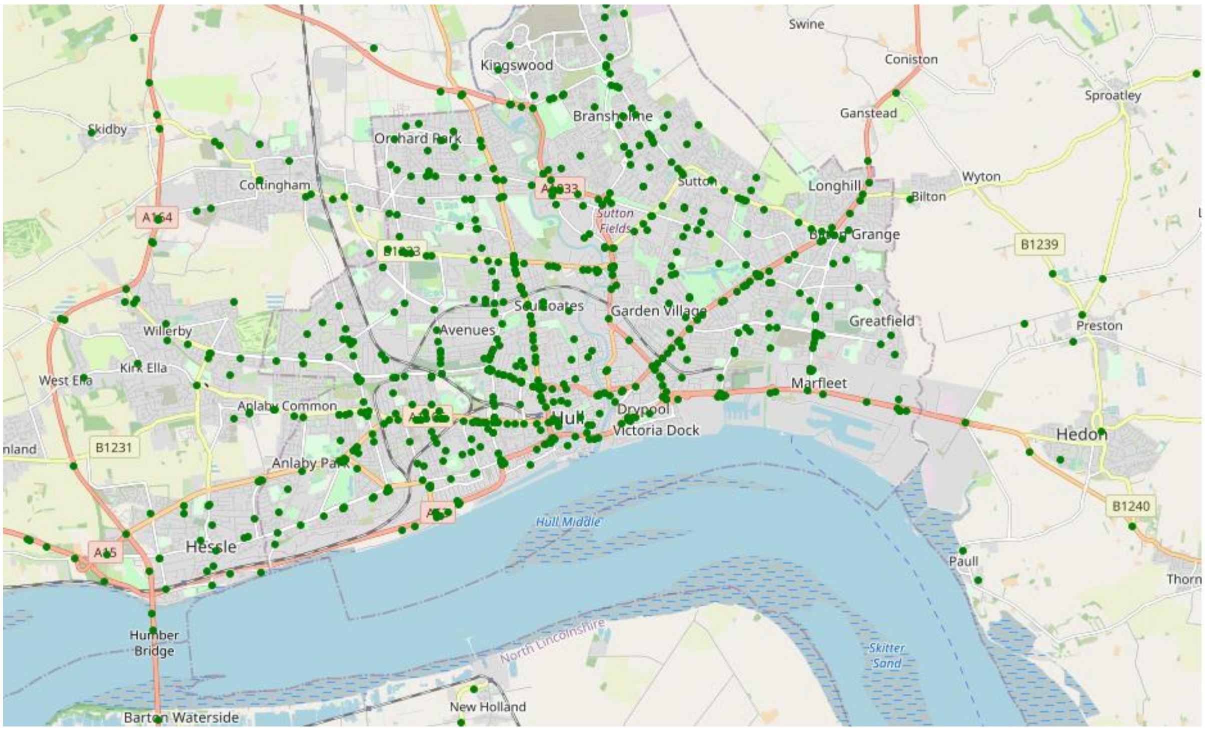


Figure 11 Folium map showing the cluster of accidents in hull for the year 2020.

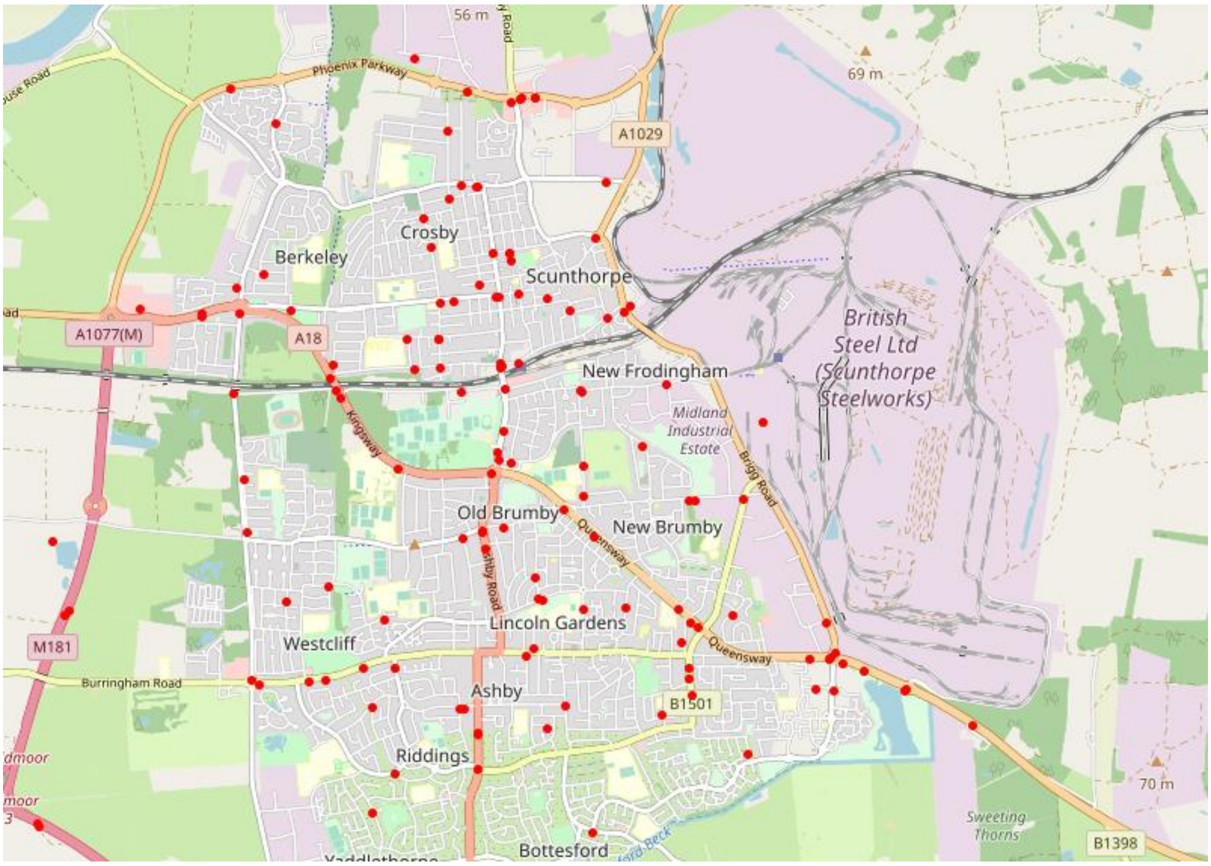


Figure 12 Folium map showing the cluster of accidents in Scunthorpe for the year 2020.

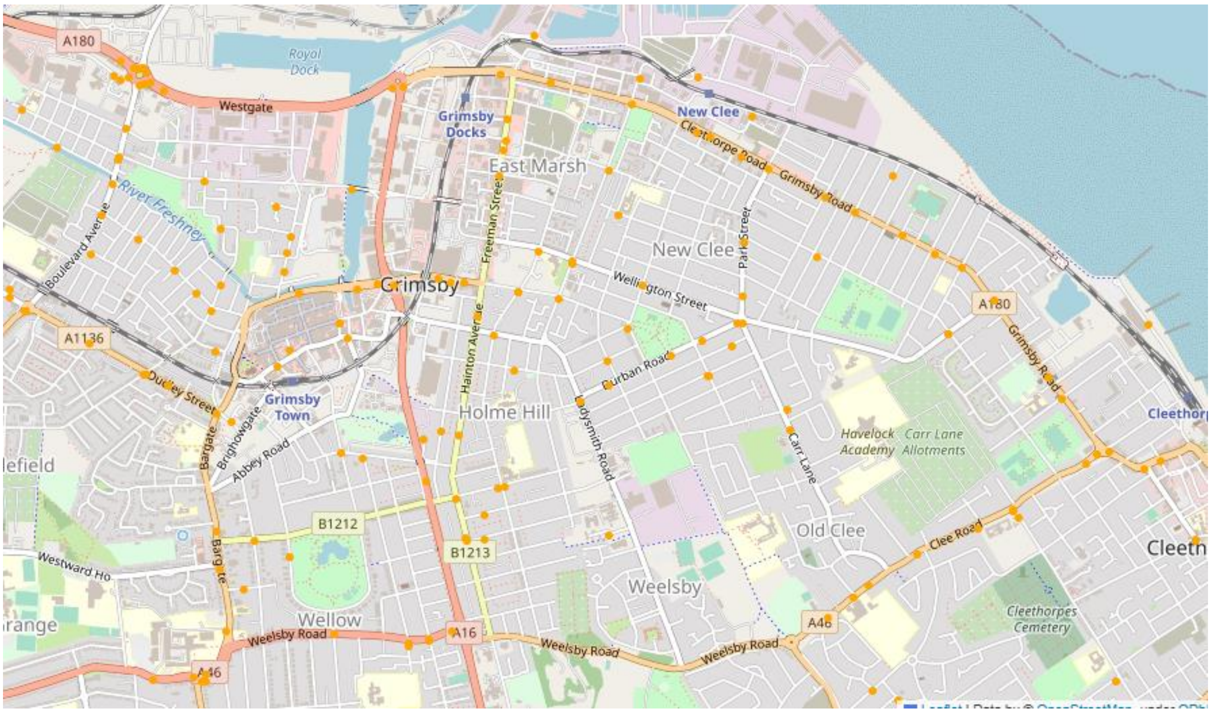


Figure 13 Folium map showing the cluster of accidents in Grimsby for the year 2020.

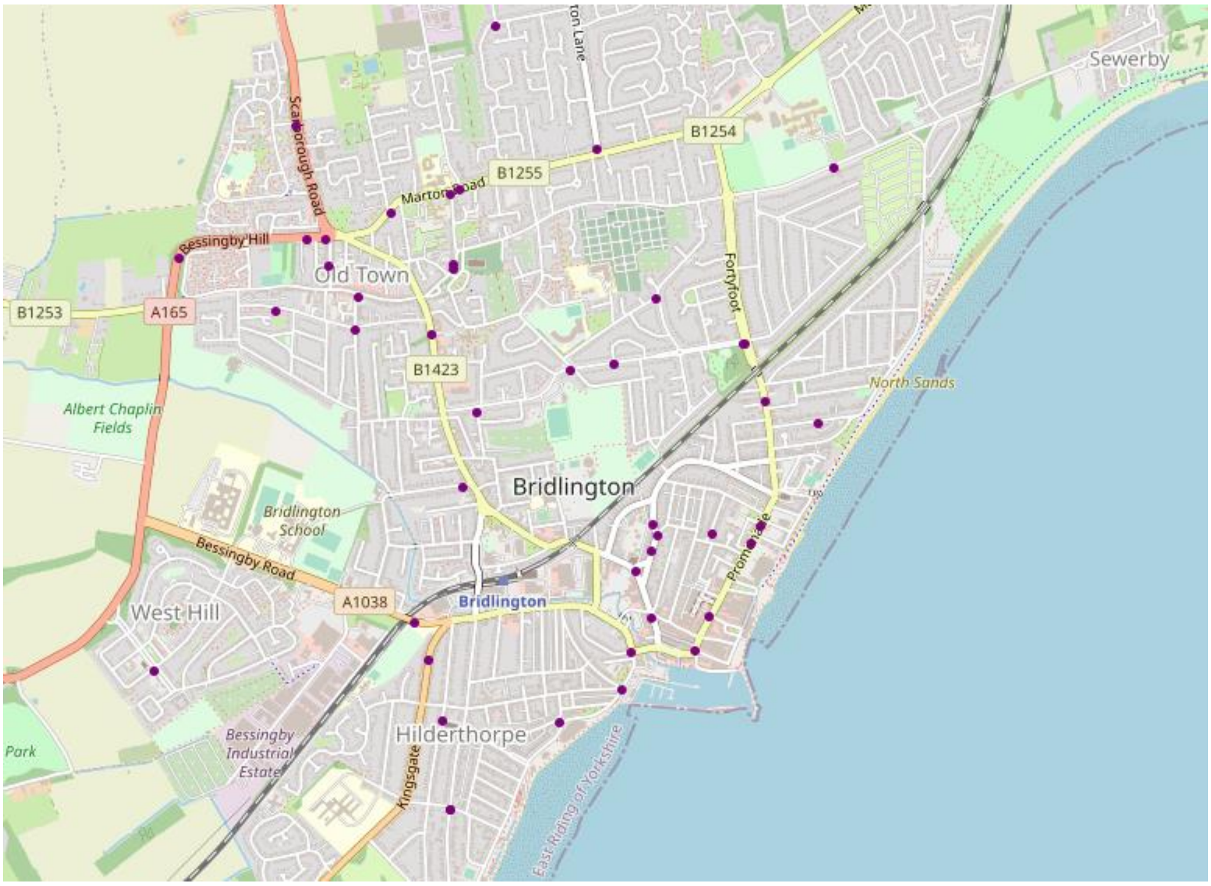


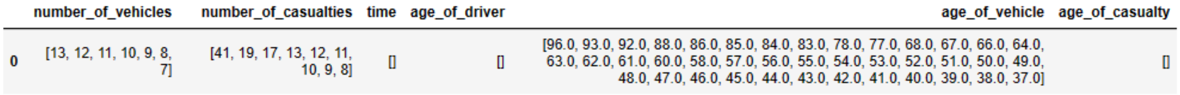
Figure 14 Folium map showing the cluster of accidents in Bridlington for the year 2020.

Outlier Detection

Before the predictive model was built, the data set was searched for the presence of outliers. To find the univariate outliers, the Grubbs test and the IQR method were used and to get the multivariate outliers, the Isolation Forest method was used.

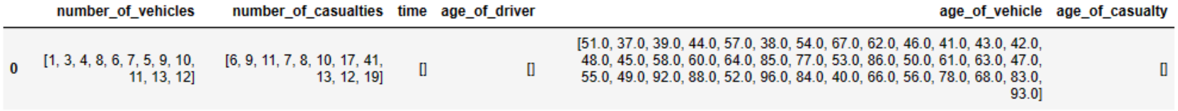
Grubbs test and IQR method

Using the Grubbs test with an alpha value of 0.05 and using the IQR method with a multiple of 3, almost the same outliers were detected as seen in Table 3 and Table 2.



	number_of_vehicles	number_of_casualties	time	age_of_driver	age_of_vehicle	age_of_casualty
0	[13, 12, 11, 10, 9, 8, 7]	[41, 19, 17, 13, 12, 11, 10, 9, 8]	[]	[]	[96.0, 93.0, 92.0, 88.0, 86.0, 85.0, 84.0, 83.0, 78.0, 77.0, 68.0, 67.0, 66.0, 64.0, 63.0, 62.0, 61.0, 60.0, 58.0, 57.0, 56.0, 55.0, 54.0, 53.0, 52.0, 51.0, 50.0, 49.0, 48.0, 47.0, 46.0, 45.0, 44.0, 43.0, 42.0, 41.0, 40.0, 39.0, 38.0, 37.0]	[]

Table 3 Univariate outliers detected by the Grubbs test using an alpha of 0.05.



	number_of_vehicles	number_of_casualties	time	age_of_driver	age_of_vehicle	age_of_casualty
0	[1, 3, 4, 8, 6, 7, 5, 9, 10, 11, 13, 12]	[6, 9, 11, 7, 8, 10, 17, 41, 13, 12, 19]	[]	[]	[51.0, 37.0, 39.0, 44.0, 57.0, 38.0, 54.0, 67.0, 62.0, 46.0, 41.0, 43.0, 42.0, 48.0, 45.0, 58.0, 60.0, 64.0, 85.0, 77.0, 53.0, 86.0, 50.0, 61.0, 63.0, 47.0, 55.0, 49.0, 92.0, 88.0, 52.0, 96.0, 84.0, 40.0, 66.0, 56.0, 78.0, 68.0, 83.0, 93.0]	[]

Table 2 Univariate outliers detected by the IQR method using a multiple of 3.

It was decided that the outliers returned from the number of vehicles and the number of casualties would be kept because although rarely, those numbers do occur, and according to Wikipedia (2021), there have been several cases across Europe involving even more casualties and vehicles than what is detected. As for the age of vehicle, it was decided that it would be ignored because the variable did not impact the results from the model.

Isolation Forest

The multivariate outliers for longitude and latitude, and location easting and location northing were detected using the isolation forest. From comparing Figure 15 and Figure 16 with Figure 17 it was seen that the locations detected by the Isolation Forest are valid locations in the UK and because the goal is to predict accident severity in the UK, these points were kept.

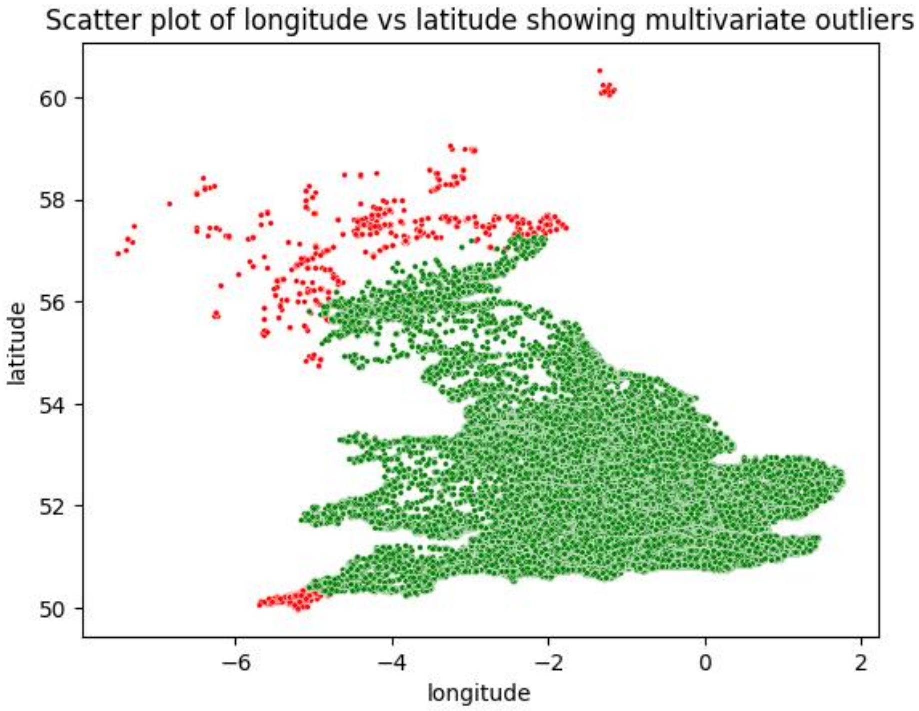


Figure 15 Scatter plot showing multivariate outliers for longitude and latitude.

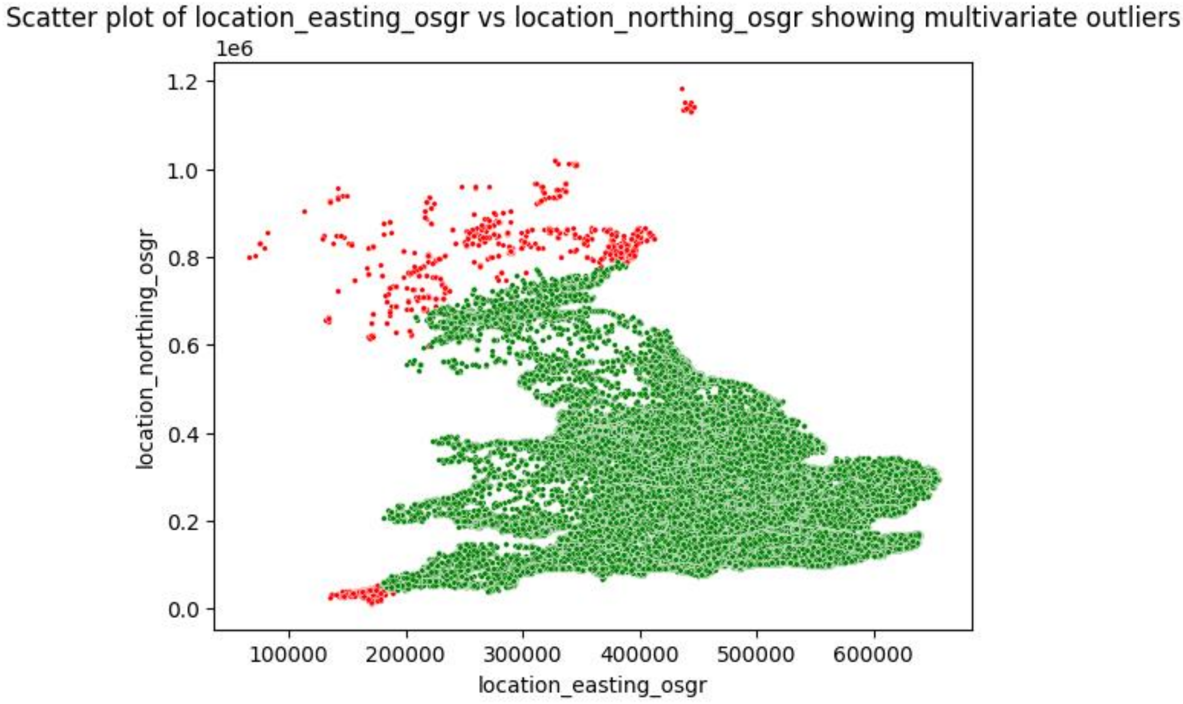


Figure 16 Scatter plot showing multivariate outliers for location northing and location easting.

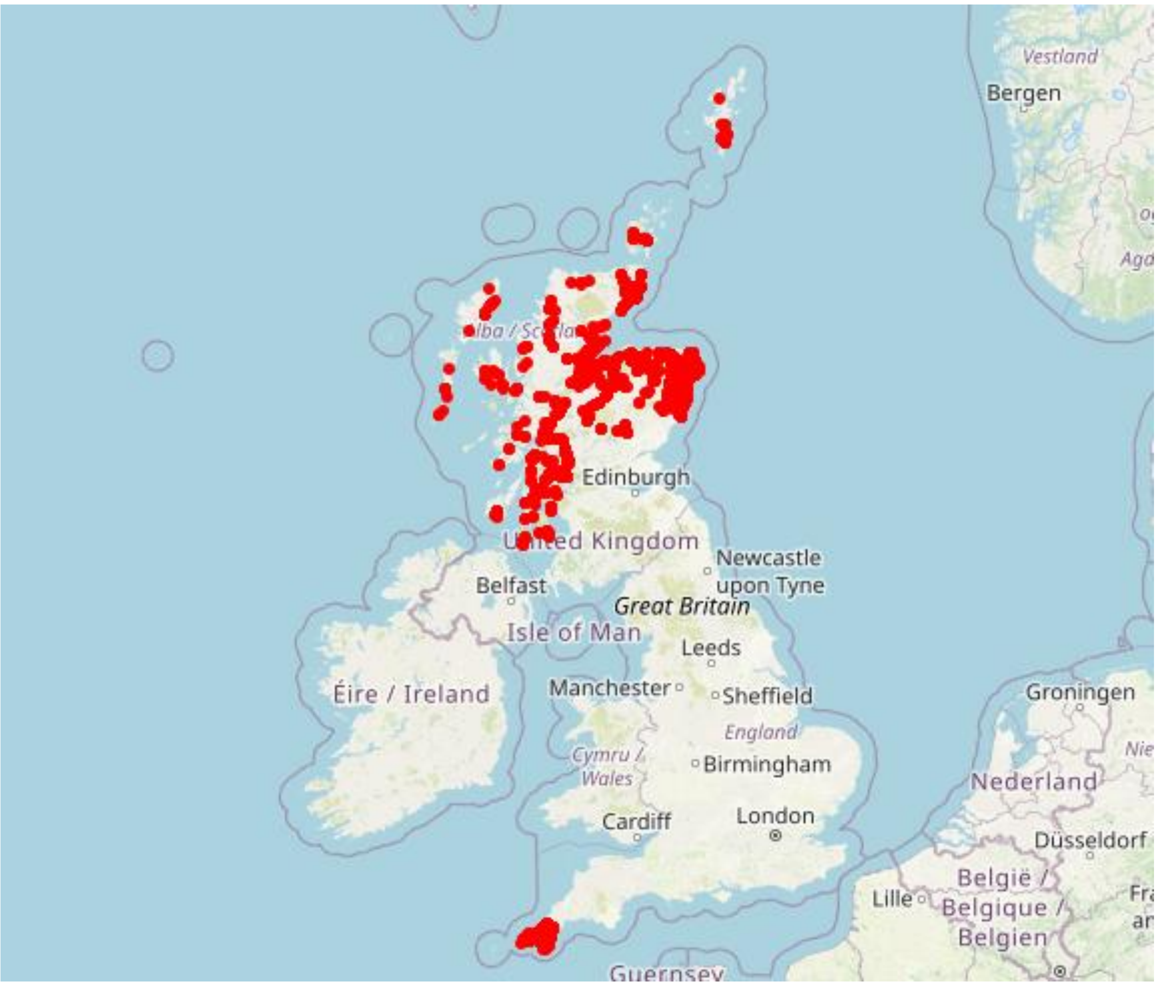
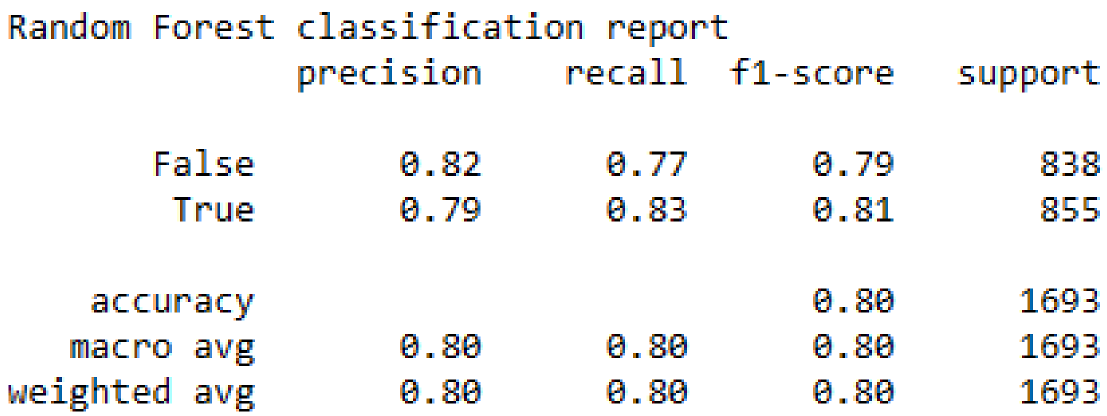


Figure 17 Folium map showing the multivariate outliers for longitude and latitude detected by the Isolation Forest on a real map.

Predictions

Three classification models were built to predict if an accident would be fatal or not which included a Random Forest, an XGBoost and a k-Nearest Neighbor model and each of them were tuned with parameters gotten from grid search in Scikit-learn (2012) to ensure that they performed best without overfitting.

By comparing Table 5, Table 4 and Table 6, it is seen that the Random Forest and the XGBoost classifiers performed best on the data set with an accuracy and recall of 80% for both and they are closely followed by the k-Nearest Neighbor which had an accuracy and recall of 77%.

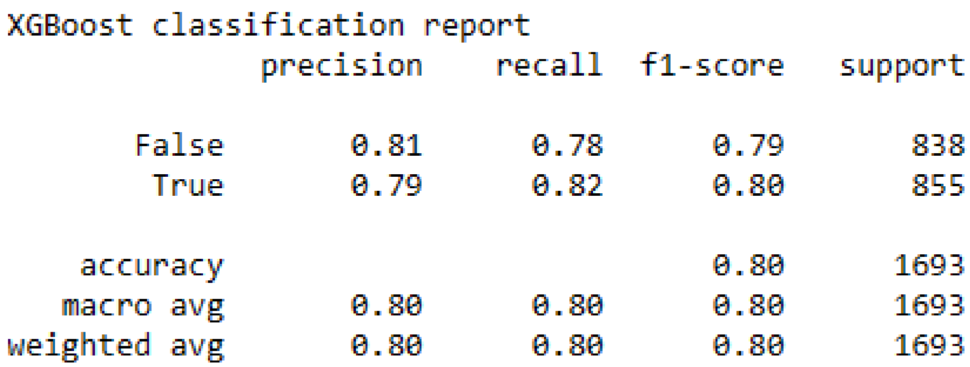


```
Random Forest classification report
              precision    recall  f1-score   support

       False       0.82      0.77      0.79       838
        True       0.79      0.83      0.81       855

    accuracy                           0.80      1693
   macro avg       0.80      0.80      0.80      1693
weighted avg       0.80      0.80      0.80      1693
```

Table 5 Classification report for the Random Forest classifier

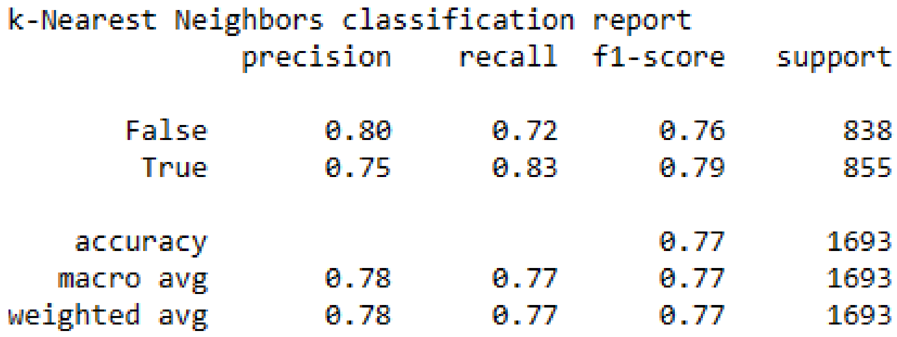


```
XGBoost classification report
              precision    recall  f1-score   support

       False       0.81      0.78      0.79       838
        True       0.79      0.82      0.80       855

    accuracy                           0.80      1693
   macro avg       0.80      0.80      0.80      1693
weighted avg       0.80      0.80      0.80      1693
```

Table 4 Classification report for the XGBoost classifier.



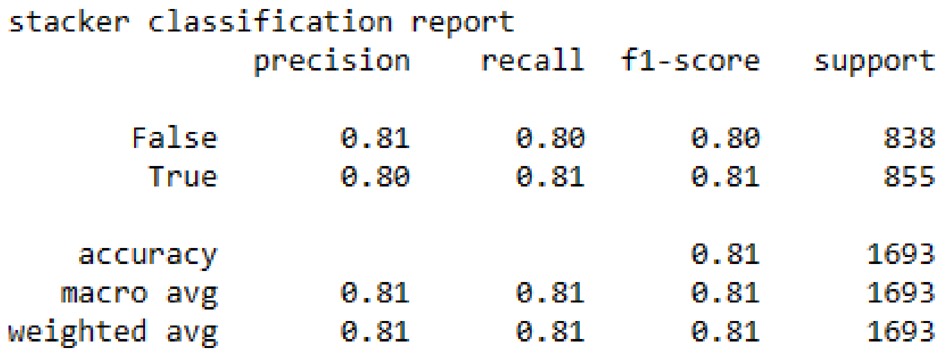
```
k-Nearest Neighbors classification report
              precision    recall  f1-score   support

       False       0.80      0.72      0.76       838
        True       0.75      0.83      0.79       855

    accuracy                           0.77      1693
   macro avg       0.78      0.77      0.77      1693
weighted avg       0.78      0.77      0.77      1693
```

Table 6 Classification report for the k-Nearest Neighbour classifier.

The three models were further stacked together which led to an improved accuracy and recall of 81% as seen in Table 7.



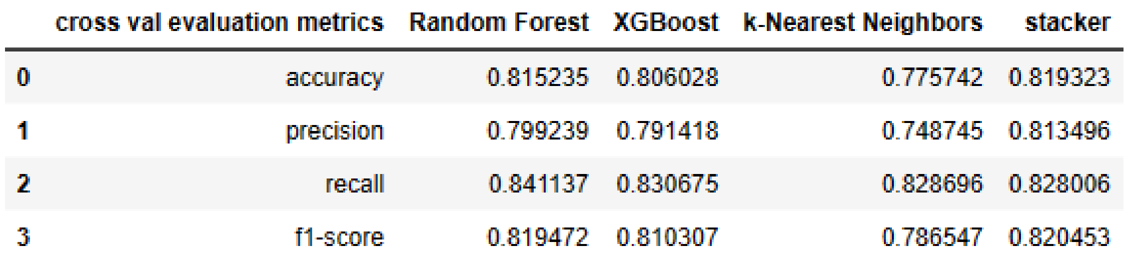
```
stacker classification report
              precision    recall  f1-score   support

       False       0.81      0.80      0.80       838
        True       0.80      0.81      0.81       855

    accuracy                           0.81      1693
   macro avg       0.81      0.81      0.81      1693
weighted avg       0.81      0.81      0.81      1693
```

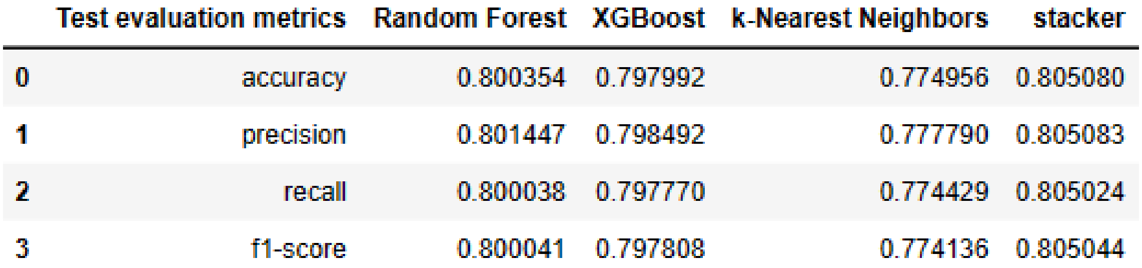
Table 7 Classification report for the stacking classifier.

As seen in the evaluations of the cross validation and the testing of all the models in Table 8 and Table 9, and also Figure 18 the stacker model performs better than every other model.



	cross val evaluation metrics	Random Forest	XGBoost	k-Nearest Neighbors	stacker
0	accuracy	0.815235	0.806028	0.775742	0.819323
1	precision	0.799239	0.791418	0.748745	0.813496
2	recall	0.841137	0.830675	0.828696	0.828006
3	f1-score	0.819472	0.810307	0.786547	0.820453

Table 8 Evaluation of the cross validation mean scores of all the models.



	Test evaluation metrics	Random Forest	XGBoost	k-Nearest Neighbors	stacker
0	accuracy	0.800354	0.797992	0.774956	0.805080
1	precision	0.801447	0.798492	0.777790	0.805083
2	recall	0.800038	0.797770	0.774429	0.805024
3	f1-score	0.800041	0.797808	0.774136	0.805044

Table 9 Evaluation of the test scores of all the models.

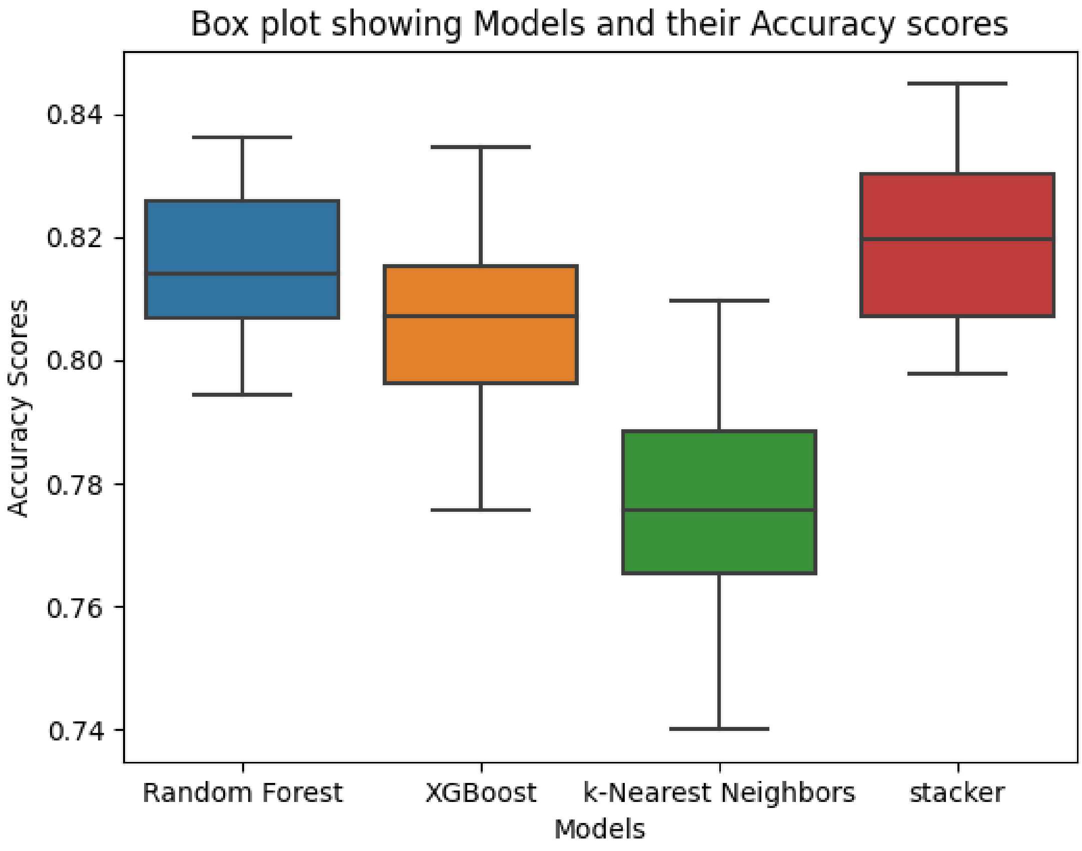


Figure 18 Box plot showing the accuracy of the different models used for classifying accident severity.

From Figure 19 which shows the most importance of variables for the classification, it is evident that the most important features are the speed limit, the urban or rural area, junction control and number of casualties.

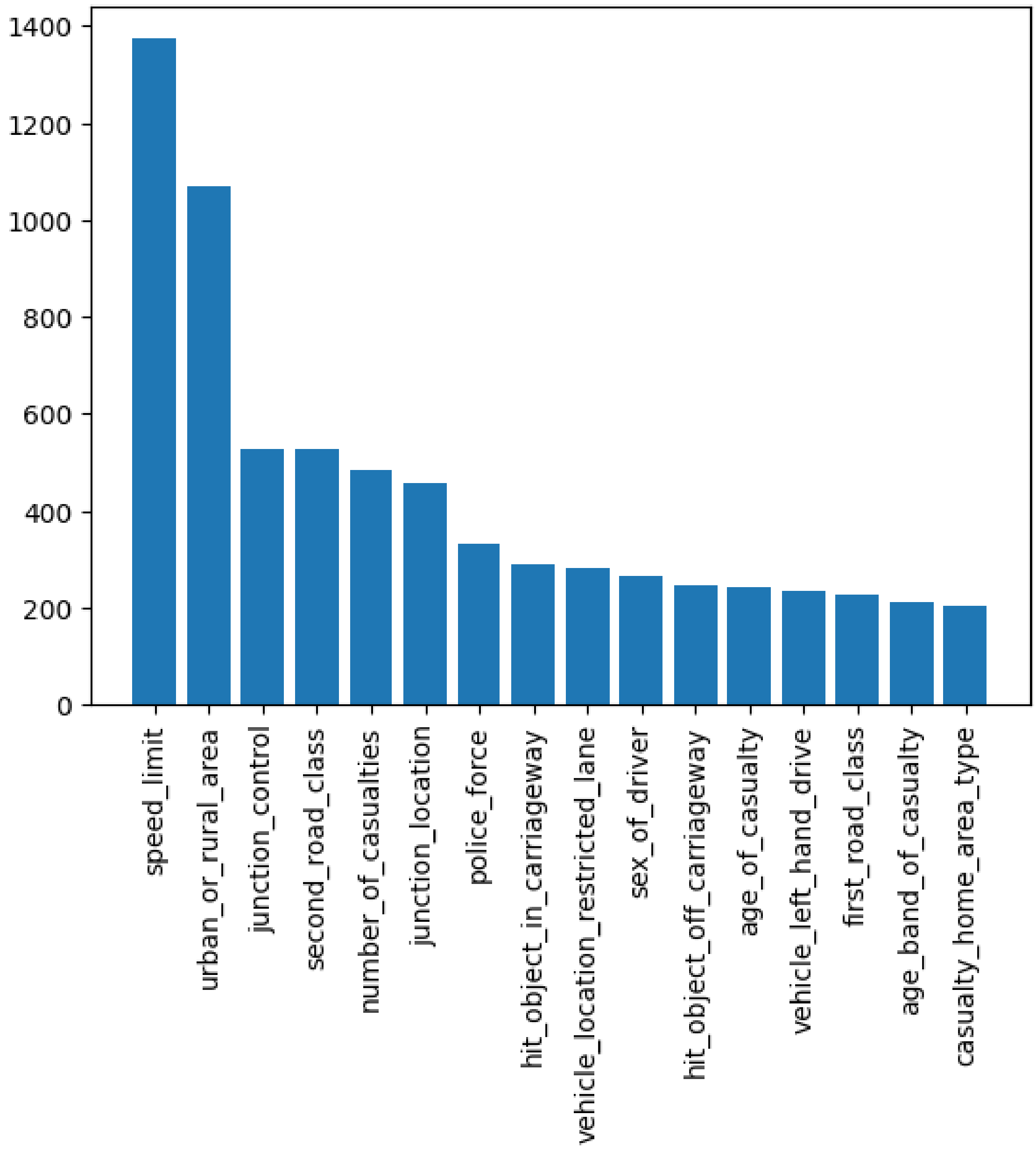


Figure 19 Bar chat showing the hierarchy of the how important different features are classifying fatal and non-fatal accidents.

Recommendations

- Traffic management should be enhanced during rush hours. This can be done by taking extra personnel or police officers to manage traffic especially on major roads.
- Based on the apriori rules, there should be consistent maintenance of road signs and symbols
- There should stress management and awareness campaigns that emphasize the importance of staying calm.

- For places with dangerous roads like Hull, measures should be taken to improve road safety. A good start would be to improve the junction controls.

Conclusion

The analysis provided insights to the when, where, and under what conditions a road safety in the UK is affected and in the end, a predictive model was built to predict accident severity with a very reasonable accuracy

Reference list

doogal.co.uk. (n.d.) www.doogal.co.uk. Available online: https://www.doogal.co.uk.

Epic (2017) *Crashes most common on Fridays*. Mackrell & Thomas Solicitors. Available online: https://www.mackrellandthomas.com/news/crashes-common-fridays/#:~:text=Gus%20Park%2C%20managing%20director%20at [Accessed 14 Aug. 2023].

Ford, G. (2022) *Hull has most dangerous roads in country with more crashes than anywhere else*. HullLive. Available online: https://www.hulldailymail.co.uk/news/hull-east-yorkshire-news/hull-most-dangerous-roads-country-6721883 [Accessed 14 Aug. 2023].

Hunter, R.F., Cleland, C.L., Busby, J., Nightingale, G., Kee, F., Williams, A.J., Kelly, P., Kelly, M.P., Milton, K., Kokka, K. and Jepson, R. (2023) Investigating the impact of a 20 miles per hour speed limit intervention on road traffic collisions, casualties, speed and volume in Belfast, UK: 3 year follow-up outcomes of a natural experiment. *J Epidemiol Community Health*, [online] 77(1), 17–25. Available online: https://jech.bmj.com/content/77/1/17.abstract.

Injury Facts (2020) *Pedestrians - Data Details*. Injury Facts. Available online: https://injuryfacts.nsc.org/motor-vehicle/road-users/pedestrians/data-details/.

More than half of car accidents happen during rush hour - Admiral.com. (n.d.) www.admiral.com. Available online: https://www.admiral.com/magazine/news/almost-half-of-accidents-happen-during-rush-hour.

Owen, J. (2019) *The Most Common Type of Car Accidents • GO-Law Solicitors*. GO-Law Solicitors. Available online: https://www.go-law.co.uk/the-most-common-type-of-car-accidents/#:~:text=Side%20Impact%20Collisions&text=These%20type%20of%20car%20accidents [Accessed 14 Aug. 2023].

Prideaux, B. and Carson, D. (2010) *Drive Tourism: Trends and Emerging Markets*. *Google Books*. Routledge. Available online: https://books.google.co.uk/books?hl=en&lr=&id=5gJ5AgAAQBAJ&oi=fnd&pg=PA146&dq=what+days+are+people+likely+to+ride+motorbikes+for+liesure&ots=s6xo0kHwvO&sig=M9_8j88ksG6rDQ0-wZmC5QbaVQ0&redir_esc=y#v=onepage&q&f=false [Accessed 14 Aug. 2023].

Scikit-learn (2012) *3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.22 documentation*. Scikit-learn.org. Available online: https://scikit-learn.org/stable/modules/grid_search.html.

Swinton Insurance (2020) *Bike Engine Types*. Swinton. Available online: https://www.swinton.co.uk/spotlight/motorbiking/guide-to-engine-types [Accessed 14 Aug. 2023].

Team, L. (2021) *What Time do Most Motorcycle Accidents Occur?* Available online: https://www.lrwlawfirm.com/what-time-do-most-motorcycle-accidents-occur/ [Accessed 14 Aug. 2023].

Wikipedia (2021) *Multiple-vehicle collision*. Wikipedia. Available online: https://en.wikipedia.org/wiki/Multiple-vehicle_collision.

Young, A. (2018) *Police reveal why there have been more crashes on A63*. HullLive. Available online: https://www.hulldailymail.co.uk/news/hull-east-yorkshire-news/police-why-more-crashes-a63-1752800 [Accessed 14 Aug. 2023].