[1.0 Introduction 3](#_Toc6509)

[2.0 Analysis 3](#_Toc6510)

[2.1 Significant Hours of the Day 3](#_Toc6511)

[2.2 Significant Days of the Week 4](#_Toc6512)

[2.3 Motorcycle Analysis 5](#_Toc6513)

[2.4 Pedestrian Analysis 6](#_Toc6514)

[2.5 Driver Age Analysis 7](#_Toc6515)

[2.6 Vehicle Age Analysis 7](#_Toc6516)

[2.7 Impact of selected variables on accident severity using the Apriori algorithm 8](#_Toc6517)

[2.7.1 Weather Conditions and Accident Severity 8](#_Toc6518)

[2.7.2 Speed Limit and Accident Severity 8](#_Toc6519)

[2.7.3 Combined Influence of Weather Conditions and Speed Limit 9](#_Toc6520)

[2.8 Regional Cluster Analysis 9](#_Toc6521)

[2.9 Outlier Analysis 12](#_Toc6522)

[3.0 Model Construction 13](#_Toc6523)

[4.0 Predictions 14](#_Toc6524)

[5.0 Recommendations 15](#_Toc6525)

[6.0 References 15](#_Toc6526)

# 1.0 Introduction

According to UK Government, 91,119 vehicle accidents take place in the UK yearly leding to approximately 1,460 deaths. Below, accident data from 2020-2023 provided by gov.uk will be evaluated. This project aims to provide insights to government agencies on how road safety can be improved. It will also, create a classification model that accurately predicts fatal injuries sustained in road traffic accidents. This will also, significantly reduce the number of deaths on UK roads. Allowing the government to formulate and implement effective road safety measures.

# 2.0 Analysis

This section will discuss the steps taken to analyse the data to gain an understanding of the results obtained from the analysis. The data located in the time column contained timestamps that required manipulation to be shown as integers. The values in the time column were converted to datetime objects in the newly created ‘converted\_time’ column. These could be converted from datetime objects in the ‘converted\_time’ column into decimals in the newly created ‘decimal\_time’ column. This made the identification of temporal trends easier.

## 2.1 Significant Hours of the Day

In order to analyse trends against various hours of the day, the time took column on the accident dataframe.

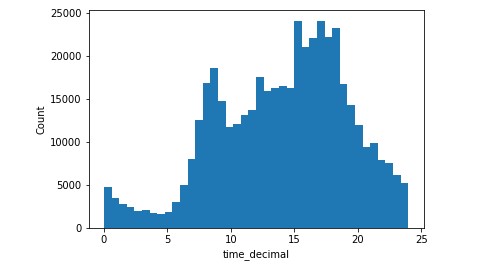


Figure 1: Amount of Accidents by Hour of the Day

Analysis of data shown in figure 1 has revealed important patterns in accident occurrences between various hours of the day. Figure 1 indicates that, between 12 pm and 7 pm, there is a notable increase in the number of accidents. Specifically, at 3 pm, a significant in accident frequency is observed, making it the period with the highest number of accidents throughout the day. The period between 3 pm and 6 pm was found to have the highest amount of accidents. This brings about the assumption that rush hour traffic and distracted drivers might be contributing factors to the peaks in figure 1. Heavier traffic volume during the afternoon hours, coupled with distractions and tiredness, increases the risk of accidents on the roads.

## 2.2 Significant Days of the Week

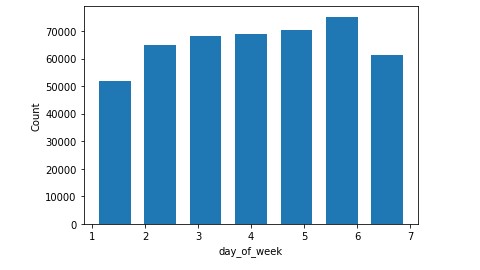


Figure 2: Amount of Accidents by Day of the Week

In figure 2, striking variations were proven in the UK traffic accident information on different days of the week. Friday stands out as the day with the highest number of accidents recorded. This finding indicates that Fridays are associated with a higher risk of accidents on the roads. Higher traffic volume, a surge in travel activities, and anticipation of the weekend may cause driver impatience or distraction. Additionally, social gatherings, events, and elevated alcohol consumption during Friday evenings might also play a role in this occurrence.

## 2.3 Motorcycle Analysis

Investigating the significant hours of the day and days of the week on which motorcycle accidents occur in the UK was undertaken and is shown in figure 3. Three categories of motorcycles: Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, and Motorcycle over 500cc. It was discovered that motorcycle accidents follow the same patterns as general accidents meaning between 12 pm and 7 pm, there is a rise in the number of accidents. Again, at 3 pm, the highest peak in accident frequency is shown below in figure 3.

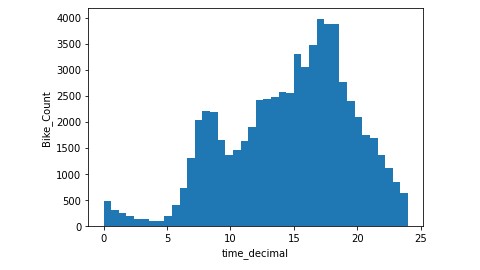


Figure 3: Amount of Bike Accidents by Hour of the Day

It is also clear that Friday is the day of the week that most accidents occurred with Thursday having a relatively high amount of accidents as shown below in figure 4.

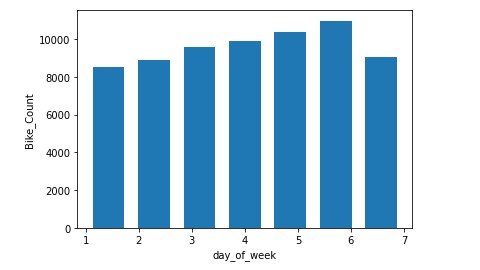


Figure 4: Amount of Bike Accidents By Day of the Week

## 2.4 Pedestrian Analysis

The pedestrian accident data in figure 5 has revealed distinct patterns in accident occurrences. It has shed light on the periods when accidents tend to be at their highest. The findings indicate that between 3 pm and 7 pm is the period that surges in pedestrian accidents peak.

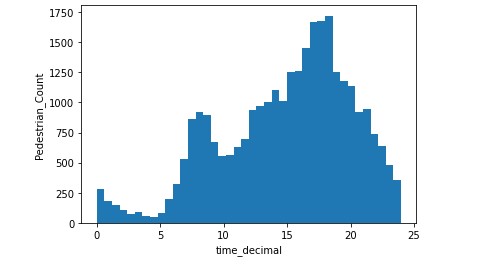


Figure 5: Amount of Pedestrian Accidents by Hour of the Day

Figure 5 reveals a notable trend in accidents based on the day of the week. It is evident that Friday stands out as the day with the highest number of pedestrian accidents. This finding suggests that Fridays are the riskiest day for pedestrians on the road.

## 2.5 Driver Age Analysis

Assessing the age of drivers presented below in 6, has provided some unmistakable results.

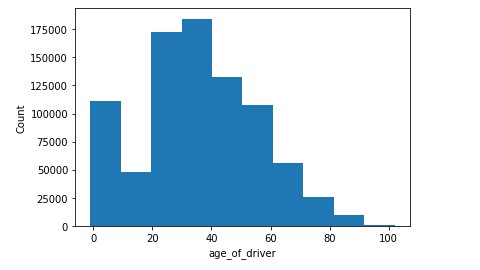


Figure 6: Amount of Accidents By Age of Driver

It can be inferred that drivers of ages 20 - 40 are most likely to be involved in accidents. The age range involved in the highest amount of accidents is between 30 and 40 with at above 175,000 people being involved in accidents

## 2.6 Vehicle Age Analysis

From observing the age of vehicles, has produced figure 7. The analysis of the age of vehicles involved in accidents has revealed significant patterns in accident occurrence. Figure 7 shows that over 500,000 vehicles between 0 - 10 years were involved in accidents.

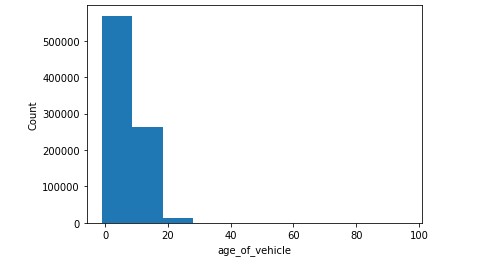


Figure 7: Amount of Accidents By Age of Vehicle

## 2.7 Impact of selected variables on accident severity using the Apriori algorithm

### 2.7.1 Weather Conditions and Accident Severity

The analysis revealed a visible association between specific weather conditions and accident severity. When weather condition, "weather\_1": (Fine weather without high winds), is present there is a high chance of accident severity being "severity\_3" (slight accident severity). The confidence value of 0.7936 indicates that approximately 79.36% of accidents involve this particular combination of weather conditions and severity.

### 2.7.2 Speed Limit and Accident Severity

The speed limit affects accident severity. The analysis found that accidents occurring in areas with a speed limit of 30 mph, ("speed\_limit\_30"), are more likely to result in "severity\_3"(slight accident severity) outcomes. The confidence value of 0.8166 indicates that there is an 81.66% chance that an accident with this speed limit will have severity\_3.

### 2.7.3 Combined Influence of Weather Conditions and Speed Limit

Further exploration of combined effects highlights additional patterns. For instance, when both

"weather\_1" and "speed\_limit\_30" are present, there is a higher likelihood of accidents of "severity\_3"(slight accident severity) occurring in areas with a speed limit of 30 mph. This association rule suggests that the combination of specific weather conditions and severity\_3 contributes to a higher probability of encountering a 30 mph speed limit area. The confidence value of 0.8132 indicates that there is an 81.32% chance of both "weather\_1" and "speed\_limit\_30" being present then an accident of severity "severity\_3" occurring.

## 2.8 Regional Cluster Analysis

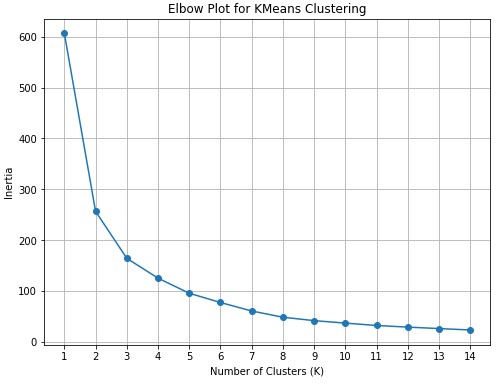


Figure 8: Elbow Plot for East Riding of Yorkshire Region

After applying the clustering algorithm, three distinct clusters of accidents were identified in Kingston-upon-Hull, Humberside, and the East Riding of Yorkshire. The value of n\_clusters = 3 was chosen because the elbow plot in figure 8 suggests that three is the best number of clusters for the data.

Clustering analysis shows the three major points where we have the highest concentration of accidents as seen in figure 9.

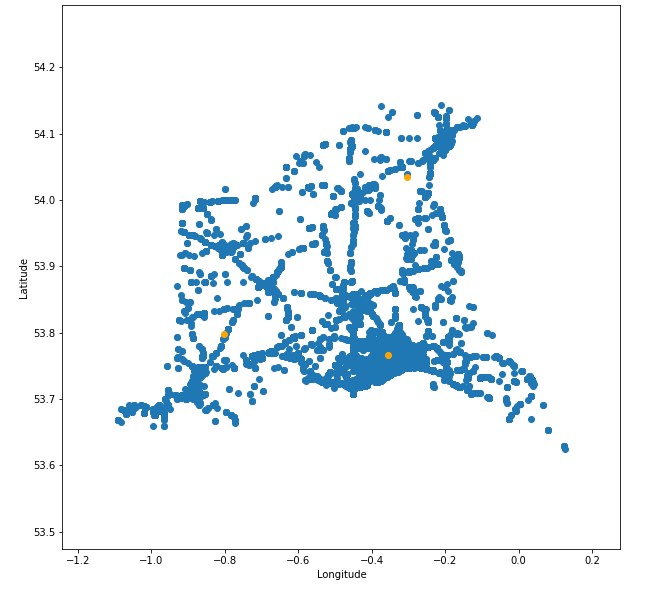


Figure 9: Cluster Diagram of East Riding of Yorkshire Region

Comparing the diagram from the cluster analysis to the map of the East Riding of Yorkshire as seen in figure 10:

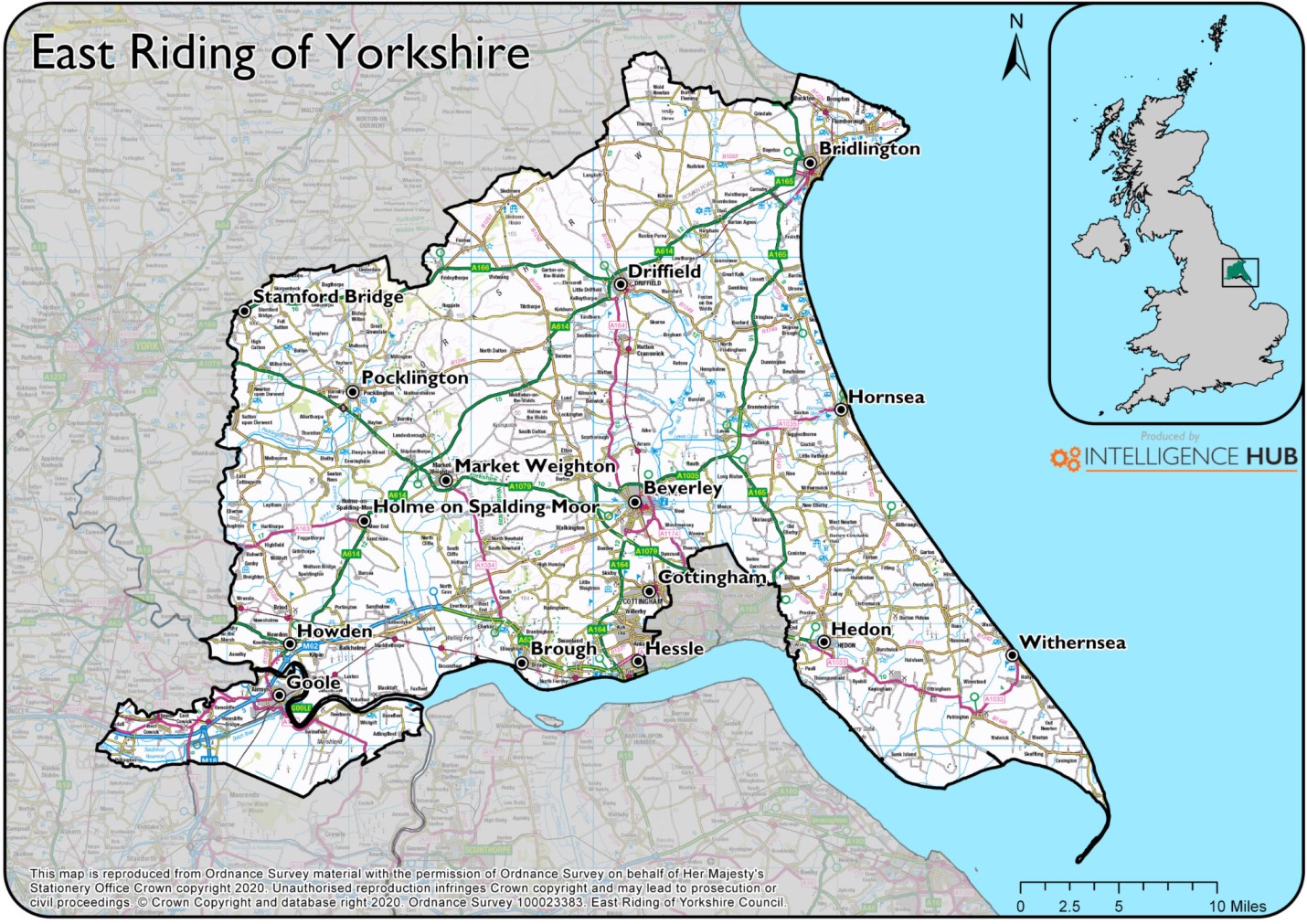


Figure 10: Map of East Riding of Yorkshire (https://intel-hub.eastriding.gov.uk/maps/)

We can see by superimposing both images that we have more accidents around Cottingham/Hull, Driffield, and Holme-on-Spalding-Moore.

Cluster 1 - Cottingham/Hull Area:

Represents an area with a high concentration of accidents, primarily centered around the Cottingham and Hull regions. The higher accident frequency in this area may be attributed to factors such as increased traffic volume, proximity to major roadways, or higher population density.

Cluster 2 - Driffield Area:

Reveals another localized hotspot of accidents around the Driffield area. The clustering of accidents in this region may be influenced by factors unique to the local geography, road conditions, or specific traffic patterns.

Cluster 3 - Holme-on-Spalding-Moore Area:

Highlights an area with an elevated number of accidents, particularly centered around Holmeon-Spalding-Moore. Further investigation into the contributing factors to this cluster may provide valuable insights into road safety measures in rural areas.

## 2.9 Outlier Analysis

Using the Interquartile Range Method on the age of the driver and the age of the vehicle for example, results obtained range from -17.5 - 90.5 years for the age of the driver and -15 - 25 years for the age of the vehicle. These numbers are modified based on domain knowledge to become 17 to 90.5 for the age of the driver and 0-25 for the age of the vehicle because, by law, the minimum age required before you can drive is 17 according to [(*Driving lessons and learning to drive*, 2015)](https://paperpile.com/c/ABeaAe/AZwU) and 0-25 for an age of vehicle because it isn’t possible to have an age that is below zero.

Using the Local Outlier Factor (LOF) method with a contamination value of ‘auto’, the following graph was generated;

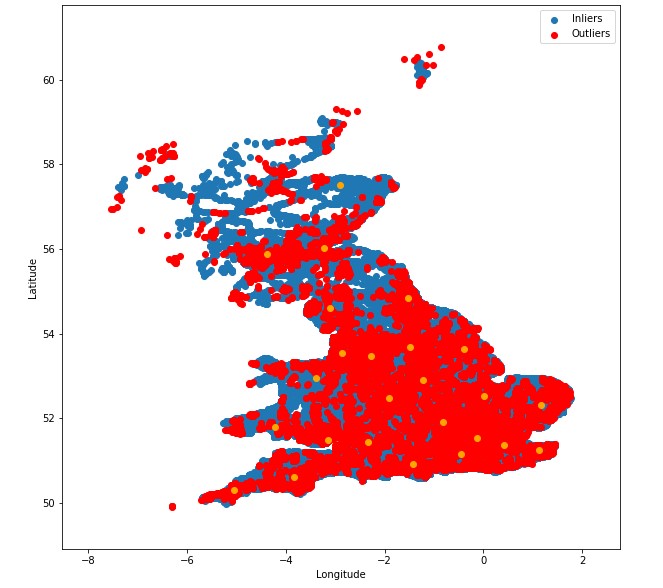


Figure 11: Outliers Using LOF

The contamination parameter determines the amount of contamination of the data set, i.e. the proportion of outliers in the data set, when fitting this is used to define the threshold on the scores of the samples [(*sklearn.neighbors.LocalOutlierFactor*).](https://paperpile.com/c/ABeaAe/Q2HL)

The outliers are marked in red and inliers are marked in blue. It is suggested the unusual entries in the dataset are kept because they seem to follow the distribution of the UK map. None of the data show any extremities.

# 3.0 Model Construction

To prepare to build the classification model, feature selection is performed using filtering and produced figure 12:

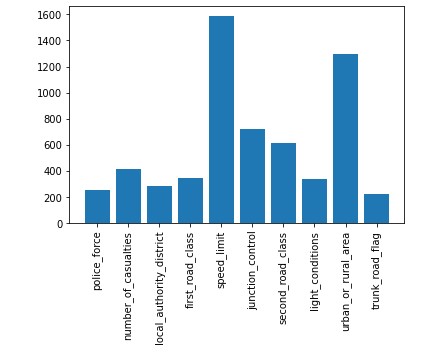


Figure 12: Feature Relevance to Classification

Figure 12 shows the importance of each feature in the classification task carried out and it is shown that speed\_limit, urban\_or\_rural\_area, junction\_control, and second\_road\_class have the best scores. A Decision Tree Classifier was used due to its ease of implementation. It is worth mentioning that a Stacking Classifier in theory would have provided higher accuracy. The data was split into test and training sets using the train\_test\_split method provided by sklearn. The test data accounted for 25% of the whole dataset.

# 4.0 Predictions

To determine the best model, a function to evaluate the best classifier model between DecisionTreeClassifier, K-Nearest Neighbor Classifier, Naive Bayes Classifier, Support Vector Machine Classifier, Logistic Regression, and Stacking Classifier is created.

**Model Evaluation**

To assess the effectiveness of our classification model, several evaluation metrics, including accuracy, precision, recall, and F1-score are employed. These metrics allow us to gauge the model's predictive performance across different aspects comprehensively.

*Accuracy*: The accuracy of the model is measured at 70%, indicating the overall proportion of correctly predicted instances among all instances.

*Precision*: Out of all instances the classifier predicted as "False ", 71% of them are actually "False." Similarly, out of all instances, the classifier predicted as "True ", 69% of them are actually "True."

*Recall*: The recall score of 70 % represents the model's capability to identify all actual fatal injuries in the dataset.

*F1-score*: The F1-score of 70% is the harmonic mean of precision and recall, offering a balanced assessment of the model's performance.

**Interpretation of Predictions**

The classification model identifies specific features that contribute significantly to predicting fatal injuries in road traffic accidents. By analyzing feature importance, insights can be gained from the factors that play a crucial role in accidents resulting in fatal injuries. These insights can assist relevant stakeholders in making informed decisions and implementing preventive measures. The features with the most importance can be seen below in Table 1:

Table 1: Table of Feature Importance

|  |  |  |
| --- | --- | --- |
| S/N | Feature | Importance |
|  | speed\_limit | 0.466582 |
|  | number\_of\_vehicles | 0.208469 |
|  | junction\_detail | 0.076166 |
|  | local\_authority\_district | 0.06687 |
|  | light\_conditions | 0.050164 |

# 5.0 Recommendations

Extensive analysis has been carried out to understand the cause of accidents in the United Kingdom. It would be suggested more sensitisation to drivers of the effect of speeding. As it is the leading cause of accidents. It is also advisable for drivers to avoid driving during harsh weather conditions. Other means of transportation should be made easily available and convenient for citizens to reduce the number of vehicles on the roads as this would lead to less traffic which would ultimately see a downward trend in the number of accidents.

# 6.0 References

[*Driving lessons and learning to drive* (2015) *GOV.UK*. Available at:](http://paperpile.com/b/ABeaAe/AZwU) [https://www.gov.uk/drivinglessons-learning-to-drive](https://www.gov.uk/driving-lessons-learning-to-drive) [(Accessed: 13 August 2023).](http://paperpile.com/b/ABeaAe/AZwU)

[*sklearn.neighbors.LocalOutlierFactor* (no date) *scikit-learn*. Available at:](http://paperpile.com/b/ABeaAe/Q2HL) [https://scikitlearn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html#:~:text=The%20 amount%20of%20contamination%20of,range%20(0%2C%200.5%5D](https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html#:~:text=The%20amount%20of%20contamination%20of,range%20(0%2C%200.5%5D.)[. (Accessed: 13 August 2023).](http://paperpile.com/b/ABeaAe/Q2HL)