**Assignment – 771762 Big Data and Data Mining PROJECT**

**Road traffic accident report**

This report explores all road traffic accidents that were reported in the United Kingdom. The data used in this project is from the government released large batch of road traffic accident data from 2019 which contains the following:

1. **Brief-guide-to road-accidents-and-safety-data**.
2. **variable lookup**.
3. **Road Safety Data - Accidents 2019**.
4. **Road Safety Data - Casualties 2019**.
5. **Road Safety Data- Vehicles 2019**.
6. **Adjustment Files**.

This report also seeks to address the following questions with detailed analyses and visualizations:

1. Are there significant hours of the day, and days of the week, on which accidents occur?
2. For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?
3. For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?
4. What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?
5. What impact, if any, do sunrise and sunset times have on road traffic accidents?
6. Are there particular types of vehicles (engine capacity, age of the vehicle, etc.) that are more frequently involved in road traffic accidents?
7. Are there particular conditions (weather, geographic location, situations) that generate more road traffic accidents?
8. How do driver-related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)?
9. Can we make predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?

The above questions will be answered with various analyses, algorithms, and visualizations.

**Analysis**

1. **Are there significant hours of the day, and days of the week, on which accidents occur?**

**Accidents occurring per day of the week:**

After analyzing the data with the aid of the diagram below, I plotted a bar plot of “Day of the week” against the “Accident index” count and I discovered that accidents occurred mostly on Fridays followed by Thursday. Friday is usually regarded as the last day of the weekdays and most people consider Fridays as a day to have some fun following the end of the stressful weekdays of working. An accident can also occur because of drunk drivers.

**Chart, bar chart

Description automatically generated**

**Accident occurring per time:**

I made a time band function to accommodate the hour range of each period of the day which is listed below as:

1. Morning Rush (6-10 am)
2. Day (10-12 pm)
3. Lunch Rush (12-2 pm)
4. Afternoon (14-16pm)
5. After Work Rush (4-6 pm)
6. Evening (6-10 pm)
7. Night (22-6 am)

After creating the time bands, I mapped them to the respective hours in which accidents occurred and I found out with the aid of the diagram below that the periods in which accident occurs mostly are in the Night (10 pm – 6 pm) and After Work Rush (4 -6 pm). This means accidents occurred mostly at night due to maybe driving home tired from work or due to bad vision at night either to human conditions or lack of streetlights.

Chart, bar chart

Description automatically generated

1. **For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?**

Motorcycles accidents during the days of the week usually occur on Friday with fine no high winds weather conditions as illustrated in the diagram below

Chart, histogram

Description automatically generated

**Motorcycle accidents by time of the day**

Motorcycles road traffic accidents occurred during the “Night” period of the day by as seen by the diagram below

Chart

Description automatically generated

1. **For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?**

**On days of the week accident involving Pedestrian**

An accident involving pedestrians during the day of the week occurred mostly on “Wednesdays" then "Friday" in daylight. The distribution plot below shows below.

Chart, histogram

Description automatically generated

**Hour of the day accident involving Pedestrian**

Accidents that involve pedestrians during the day under Speed Limit conditions. From my analysis and visualization, it was discovered that accident that involves pedestrian occurs mostly at “Night” with a speed limit of “30mph”.

**Chart, histogram

Description automatically generated**

1. **What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?**

According to **Timeanddate.com. 2022. *Daylight Saving Time Changes 2019,*** the Daylight-Saving time started on the 31st of March 2019 and ends 27th of October 2019. Therefore, I made a count plot of the weeks before and after the daylight savings started and ended, daylight savings started on the 13th week in 2019. The diagram below shows the three (3) weeks between and after the daylight started, I discovered there was a significant impact on road accidents as the previous week before daylight savings started the accident count was 2092, while the week when the daylight savings started there was 2264 accident count that’s about 172 road traffic accident increase.**Chart, bar chart

Description automatically generated**

Daylight savings ended on the 43rd week in 2019, according to my analysis with the aid of the diagram below, I discovered that the week daylight savings ended road traffic accidents reduced by a difference of 238 as the week before it ended the accident count was 2446 while the week daylight savings ended it was 2208. Therefore, daylight saving ending didn’t really have a significant impact on road accident.

Chart, bar chart

Description automatically generated

1. **What impact, if any, do sunrise and sunset times have on road traffic accidents?**

Hypothetically, in most cases the sun rises by 6:00am a duration lasts for 3 minutes and starts to set by 6pm. Therefore, I analysed the Time in hour by getting all the hours which accident occur after 6am and after 6pm. I discovered that sunrise had a 1.94% impact on accident which wasn’t significant, and that sunset had a 24.83% impact on road traffic accident.

1. **Are there particular types of vehicles (engine capacity, age of the vehicle, etc.) that are more frequently involved in road traffic accidents?**

To examine if Conditions like types of Vehicles “Engine\_capacity” and “Age\_of\_Vehicle” have effects on road traffic accidents, the Kmeans Clustering algorithm was used to determine that. The number of “K” clusters to use was determined by the elbow method function, after plotting the scatterplot, I discovered that clusters were formed around the 5000 and 10,000 Engine Capacity with 0-10years old age of Vehicles.

Chart, scatter chart

Description automatically generated

1. **Are there particular conditions (weather, geographic location, situations) that generate more road traffic accidents?**

To determine if conditions like weather, geographic locations, and other features contribute to more road traffic accidents. To accomplish this, I used K-Means Clustering, which is a type of unsupervised learning algorithm, density based clustering algorithm, and used Associate Pattern Mining on some of the selected features of the dataset.

**For K-Means Clustering**: I used speed limit and Weather conditions as my features then used the elbow method to find the number of "K" to be used. After plotting the scatterplot, I discovered most of the clusters are around 2 which represents rain under speed limits of 30mph and 60mph. This means accident occurs mostly when it's raining with no winds at 30mph and 60mph. The diagram below illustrates it:

A picture containing different

Description automatically generated

**Density-based clustering (DBSCAN)**

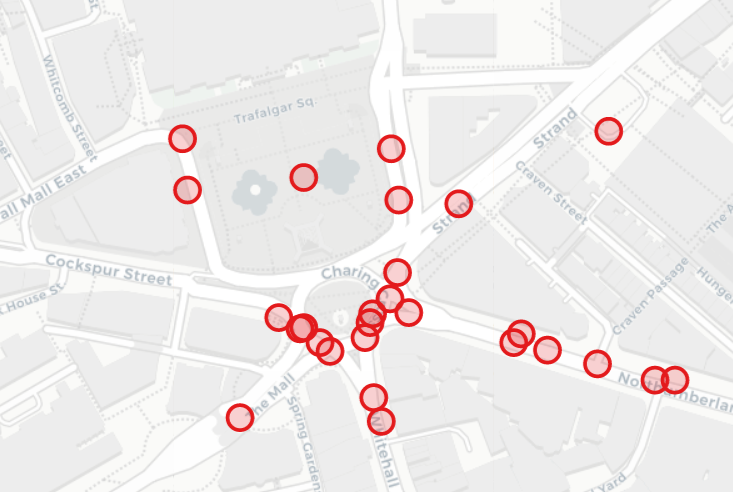
For further analysis of the locations of accident hotspots, I used a Density-based clustering algorithm to identify the accident hotspot area of two locations which were the two most occurring locations using the **“value\_counts()”** function.

The first location was in the city of London while the second was the Westminster area. The picture below shows where the major accident hotspots are between both cities.

Map

Description automatically generated

After zooming the red cluster areas where accidents occurred, I discovered that accidents occurred majorly at the charing roundabout as shown below:



**For Associate Pattern Mining:**

Associate Pattern is generally used for identifying different variables in datasets. For the Associate Pattern mining, I used the apriori algorithm to identify the association rules between three features in the datasets which are “Accident Severity”, “Speed Limit” and “Weather Conditions”. Generally, the Apriori algorithm has three important factors which are support, confidence, and lift. These factors were implemented on the above features to understand the different associations between the three features.

Graphical user interface

Description automatically generated

Using the index 3 value as an example on the diagram above to explain the associate pattern:

1. The probability of Weather\_Condition\_1(Fine no high winds) causing accidents is about 78.5% while the probability of Severity\_3(Slight) injury in road traffic accidents is 78.6%.
2. The support for weather conditions (Fine no high winds) and accident severity(slight) is 61%
3. The correlation between weather conditions (Fine no high winds) and accident severity(slight) is 0.97

Hence, we can say weather conditions (Fine no high winds) and accident severity(slight) as an impact on road traffic accidents.

1. **How do driver-related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)?**

Using the SelectKBest method, I used Age of Driver, Journey Purpose, Sex of Driver, Age of Vehicle, and Day of Week as features to predict the impacts they have on a traffic accident. I discovered that the Age of the driver had a significant impact on road accidents followed by the sex of the driver as illustrated below.

Chart

Description automatically generated

**Prediction**

**Can we make predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?**

To make a prediction I used various models on selected features with the highest correlation to accident severity which are listed below to get a good accuracy score:

***Accident\_Severity, Casualty\_Severity', Police Officer Attend, Sex\_of\_Driver,2nd\_Road\_Class, Location\_Easting\_OSGR,***

***Longitude, Junction\_Contro, Weather, Casualty\_Type, Junction\_Detail,2nd\_Road\_Number,***

***Vehicle Propulsion Code, Ped Cross - Physical,Journey Purpose,1st\_Road\_Class, Vehicle\_Type, Number\_of\_Vehicles,Engine\_Capacity\_(CC), Day\_of Week, Junction\_Location, 1st\_Point\_of\_Impact.***

I created an X and y variable to cater to the features and target (Accident Severity) to be able to train my datasets properly.

The train\_test\_split feature of the sklearn module in python was used to split the datasets to train and test with a ratio of 70% train and 30% test.

After using DecisionTreeClassifier I was able to get an accuracy score of 97% which was a decent accuracy. A table showing different classifiers used to predict the severity of accidents is shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | DecisionTreeClassifier | Gaussian Naïve Bayes classifier | Random Forest Classifier | After Stacking multiple Models |
| Accuracy score | 97% | 82% | 97% | 97.5% |

**Recommendations**

After analyzing the data of Accident, Vehicle, and Casualties datasets extensively with numerous Visualisations and algorithms these are my recommendations that I would suggest to the Government to reduce road traffic accidents and ensure safety:

1. There should be compulsory enforcement of proper kits which include helmets, front light/backlight and a High-Visibility vest must be worn by all bicycle users cycling on the road.
2. Ensuring that money collected as road tax is fully utilized for continuous maintenance and construction of roads.
3. There should be more Zebra crossings provided for pedestrians at suitable places to make roads more secure for pedestrians.
4. There should be mandatory road accident/safety topics and projects in the curriculum for children to learn at an early age to be well informed about safety measures and road accident prevention methods to reduce road traffic accidents.
5. Existing traffic rules should be properly enforced as it has been seen that one of the reasons why traffic violations as increased are because rules and regulations have not been properly imposed.
6. Children below a certain age should not be allowed to cycle on busy roads or roads where heavy-duty vehicles are plying.

**Reference**

**Timeanddate.com. 2022. *Daylight Saving Time Changes 2019 in London, England, United Kingdom*. [online] Available at: <https://www.timeanddate.com/time/change/uk/london?year=2019> [Accessed 28 April 2022].**

**National Radio Astronomy Observatory. 2022. *Do We Observe the Sunrise Six Minutes After it Actually Occurs? - National Radio Astronomy Observatory*. [online] Available at:** **<https://public.nrao.edu/ask/do-we-observe-the-sunrise-six-minutes-after-it-actually-occurs/> [Accessed 29 April 2022].**