

COS7048-B MSc Group Project

Coursework – Report 1

Team 3

Ali Mohammed – 20002497

Eniga Ahiante - 20026427

Ethan Daniel – 17022218

Funmilayo Celestina Ayeni - 19022115

Saba Rasheed - 20029056

Saif Javed - 16007638

Abstract

*With the rise of the digital economy, more data is being captured in various fields such as health, entertainment, and education. Data systems provide a solution for datasets that portrait. Data analytics is a core component within this subject as this is the main purpose of capturing data to begin with. This paper will detail machine learning systems and experimenting in line with a problem statement with the goal of providing a detailed analysis and a proposed solution. Road traffic safety is an essential facet in every nation as it keeps people safe and precludes damages and deaths. Whilst it is difficult to avoid all accidents through road safety procedures, road traffic accident data can be examined to deliver a plan of where upgrades to roads need to be made.*

Table of Contents

[Introduction 3](#_Toc66975198)

[Problem Statement 3](#_Toc66975199)

[Aims and objectives. 4](#_Toc66975200)

[Issues and Risks 4](#_Toc66975201)

[Requirements 5](#_Toc66975202)

[Literature review and research 6](#_Toc66975203)

[2.1 Data sources 7](#_Toc66975204)

[Methodology 8](#_Toc66975205)

[Dataset 8](#_Toc66975206)

[Output Class Analysis 10](#_Toc66975207)

[Module Experiments and Data Exploration 10](#_Toc66975208)

[Initial prototype / results (Results and Discussions) 11](#_Toc66975209)

[Future work/reflection (Conclusions) 16](#_Toc66975210)

[Bibliography 17](#_Toc66975211)

[Peer evaluation 18](#_Toc66975212)

**GitHub Repository:** [**https://github.com/UoBGroup3/GroupProject**](https://github.com/UoBGroup3/GroupProject)

List of Figures and Tables

Figure 1: ……………………………………................................................................................................................5

Figure 2: ……………………………………................................................................................................................8

Figure 3: ……………………………………................................................................................................................10

Figure 4: ……………………………………................................................................................................................11

Table 1: ……………………………………………………………………………………………………………………………………………....8-9

Table 2: ………………………………………………………………………………………………………………………………………………….11

Table 3: …………………………………………………………………………………………………………………………………………….11-12

# Introduction

Road traffic wellbeing is a significant angle in each country as it keeps residents safe and forestalls wounds also, fatalities. While it is difficult to forestall all mishaps through street wellbeing measures, street traffic mishap information can be breaking down to give a "map" of where enhancements to streets should be made. The Imperial Society for the Anticipation of Mishaps (RoSPA) in the Assembled Realm have expressed on their site that "the most recent couple of many years have exhibited that viable and far-reaching street security procedures can diminish the quantity of individuals killed or harmed out, regardless of expanding traffic levels.

England presently has outstanding amongst other street wellbeing records on the planet - yet with around five individuals actually dying on England's streets consistently there is substantially more to be finished." The department for Transport UK (United Kingdom) (DfT) are an administration office that deal with the vehicle organization in the Unified Realm. DfT distribute a dataset two times per year on auto collisions in the UK. They have distributed six vital targets in 2019 which mean to convey a superior support of general society. While nearby committees work on street security; it is likewise a critical factor to consider for DfT as they need to guarantee the street traffic network is inacceptable stream and this will not be conceivable without protected and dependable streets for the general population.

# Problem Statement

Accidents generally are unpredictable, and, in this task, we will be working with the public dataset by DfT on the accidents in Great Britain. This is to determine the probability of an accident occurring under certain conditions using the data set. An interactive approach is being used to provide critical analysis to satisfy the business requirements.

After reviewing the data with various functions, questions such as below were raised:

* Probability of having an accident in specific cities or locations. (using classifications such as Low, Medium, and High).
* Make a model predication using a future trend.
* Deduce human, road, weather, and vehicle factor.
* Effect of mental health of the driver, poor road geometry or vehicle condition.
* Accidents due to human behavior and error.
* Unidentified road environment factors that can cause accidents.
* Effect of operational policies such as speed limits, driving ambience and traffic.

Here the data set from 1979 to 2019 have been combined into one larger data set to give a more precise assessment of the huge information framework. The information comprises of several millions of accidents which have occurred in Great Britain (data.gov.uk, 2020).

# Aims and objectives.

Our aims for this project are to increase and improve the measures of effectiveness of road safety Education, and Publicity projects to the public.

* Improve the public basic road awareness thus providing a base on which to build future skills.
* Develop safe behaviour.
* To advise the public of their awareness, skills, knowledge, and behaviour of road safety accidents.
* To increase knowledge of road traffic accidents in the public by providing relevant information.

To help protect all road users, as well as portraying this information to the public we need to understand what we can put in place to protect them. through information and results for safer road users, safer roads and infrastructure or technological innovation for safer vehicles. Changes in behaviour are most effective when they are progressive; we need to be receptive to current ideas, reflect, plan, act and maintain the behaviour to make real change.

Our objectives are in this project to study the triggers of accidents and suggest corrective actions at prospective location. To assess existing design. To compute and show the public the exact time and area with selected features on where accidents occur mostly, therefor making the public more cautious.

* To increase driver awareness of standards in relation to road safety.
* Use data set from 1979 to 2019 to predict road safety accidents in specific areas.
* To use the selected data to show live time and date, data prediction modelling showing to the public.
* Reinforcing Public Awareness in road safety.

Road safety is especially important nowadays as there has been too much of increase in vehicles on road and it becomes crucial that we follow the traffic rules and law and save ourselves and our families from road accidents. Also making other people aware of road safety and aware of the accidents in specific area.

# Issues and Risks

There are, “human activities that have ethical issues in need of investigation and systematic discussion” (Hanson, 2017, p. 4). Traffic is just one of these. Use of personal vehicles has always been a cause of harm in modern world according to Husak (2004, p. 351). According to the World Health Organization (WHO, 2015), 1.2 Million people are killed annually in road crashes. However, fatalities are not the only issue when discussing moral issues of traffic and safety. Examples of general aspects of consideration are “infrastructure” and “transport systems.”

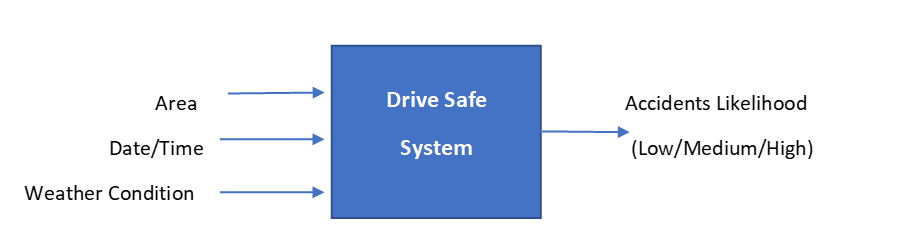
Ethics deals with issues of rights, obligations, duties, and responsibilities. It also elaborates on whether the intent of these theories and discussions and the definition of these activities are right or wrong. Traffic ethics have turned the knowledge of the essence, strength, positive and negative sides of traffic into a moral value system and code that will provide humans proper, careful, useful, and legal participation in traffic and its uses.

One of the issues of dictating traffic laws is that it must be clear on the functions of everyone for his or her responsibility in traffic. There is a constant need to warn people of the negative moral procedures and for everyone to control themselves to prevent their behaviours to cause suffering to themselves or others.

They are many risks concluding as time goes on, they are more growth of cars on the road which increases the probability of accidents to increase as more car users are on the road. Passage of people and commodities on the road is essential for societal, economic, and constitutional grounds, but this needs to travel guides to a risk of road traffic damages. A variety of factors establish who uses different components of the transport system, by what means they use them and wherefore, and at what point in time. It may not be viable in put into practice to eradicate all risk, but it is conceivable to reduce some.

# Requirements

1. Deploy a road safety system that uses the publicly provided dataset by the Department for Transport (DfT) to deliver some useful knowledge to the public.
2. The system should be able to predict the likelihood of accidents in specific local authority area in Great Britain given date, time, and weather conditions as figure 1 shows.



*Fig 1. Data flow diagram*

1. The overall probability of accident in specific area regardless the weather and time information should not be a public information as we aimed to market this system feature to the car insurance companies as customers.
2. The system interface should be simple and show interactive charts and dashboards by default about the areas with highest accidents likelihood in the current browsing time.
3. The Accuracy of the system should be acceptable with more than 90% accuracy for the high likelihood class.
4. The system should be scalable and adjustable to learn from the new provided accidents data.
5. The system Performance and responsive time.
6. Availability of data to the public and ensuring redundancy and security of data for future references.
7. Formulating the data in such a way that it can be easily visualized and analysed to promote intelligent business decision making.

# Literature review and research

In urban areas, motorways, and other UK roads, road safety has always been a serious problem by itself and affects the capacity and performance of the road network. In 2020, around 1,580 people were killed in roads accidents in the UK alone. The technology for roads and cars is constantly improving. Today's motorways have several features such as variable speed limits, warning and information signals, smart motorways, SOS phone boxes, congestion monitoring, and digital speed cameras. New vehicles are being developed for easier and safer driving. key factors such as vehicle age, safety measures, human blunder, time, and a spot of misfortune are major contributors to road accidents.

Various authors have analysed different cause and effects of road accidents on the everyday lives of individuals involved and the economy some of these effects are highlighted below:

Effects of Weather on Road Accidents

It is said weather affects mobility; however, these effects differ on the type of roads, due to this the number of causalities and injuries differ depending on the effect of mobility (Hayat et.al., 2013). Weather has a huge effect on mobility, because mobility is enhanced through friction and a slight change in one of the two surfaces crossing at a particular moment can cause a slip. Therefore, during a severe weather (e.g., heavy snow or rain) the number of accidents and casualties recorded during the accidents is at its peak (fhwa.dot.gov, 2021). During this season caution is advised by the DfT to citizens when using the roads. The papers reviewed in this research work analysed the added risk of driving during a severe weather, Harold Brodsky in 1987 used two statistical methods to analyse the risk factor and his results were a 91% possibility of having an accident during a severe weather.

Effects of Road Accidents on the National Economy

How much exactly does road accidents cost our economy the factor taken into consideration is the economy’s Gross Domestic Product (GDP), irrespective of how developed a country is, it is estimated by world bank that at least 2% of its GDP is used in developing a road accident prevention and management plan (worldbank.org, 2021), and it also estimates that countries that do not invest towards road accident prevention and management may lose 22% of their GDP. The criteria to estimate the cost of road accidents should be as follows:

* The cost of estimate should not be too old,
* The estimate should include all road accidents.
* The estimate should include direct and indirect cost of accidents, including an estimated cost for life.
* It should also include the GDP of the country for the year (Rune Elvik, 2000).

Effects of Speed on Road Accidents

Rise in technology as seen the introduction of state-of-the-art cars that can move very fast, many young people however ignore the dangers of over-speeding and simply ride for the thrill it brings, according to the DfT over 20% of road accidents record is caused by over-speeding. The higher the speed the more likelihood an accident will occur (Taylor et.al., 2000), a 5% change in accidents was mostly because by a 1 mile/hour change in speed. Taylor et.al., in their research follow two main approaches which is driver-based and road-based, in the road-based they focused on discovering what type of roads accidents caused by over speeding where more prominent on and what roads and were they were not, and in the driver-based they focused on the state of mind of the driver, if driving under the influence of alcohol and drugs or not cause the increase in speed? The results were that in each set of traffic and road condition the frequency of accidents increases basically with the speed of the traffic and vice-versa.

Effect of Drunk Driving/ Tiredness on Road Accidents

“Tiredness is insufficiently recognised as a cause of road accidents” (Brown, 1994), he performed an extensive literature review and is conclusion was as stated above, he also stated that the main effect is a “progressive withdrawal from road and traffic demands”. In 1980 a statistical research by Narim shows that 1% of road accidents casualties are caused by tiredness and 48% of its fatalities. In a questionnaire created by Fridulv Sarberg in 1999 during his research into drowsiness as a cause of road accident he discovered that crossing the edge line on the road while feeling drowsy was a major reason for accidents as it happened 40% of the time, he also discovered that age also influenced road accidents as 50% of road accidents caused by tiredness were caused by drivers of the older generation.

Driving under the influence is frowned upon because in that state the sense of sight is skewed (swan, 2015), and it has contributed to a huge number of road accidents and casualties. In a survey conducted by the NHS in 2016, 17 % of reported accidents were cause by intoxicated drivers, 13% of which resulted in the deaths of the accident victims and 9% of which suffered from serious injuries. The legal limit for alcohol level of drivers is 8o milligrams of alcohol per 100 millilitres of blood, 35 micrograms per 100 millilitres of breath, or 107 milligrams per 100 millilitres of urine (drinkaware.co.uk, 2021).

DfT has identified many factors related to road traffic accidents. Certain factors have now been well identified but not limited to, effects on the road surface, weather, driver experience, knowledge of road signs, and age of the driver.

## **Data sources**

The improvement of technology has seen road safety become an important part of our everyday lives, as research into the various causes of road accidents and the importance on educating the public on road safety measures take precedence a lot of data has been recorded and the data used in the research is on road accidents in the UK, the data investigates the various roads and motorways present in the UK and compares the varied factors that causes road accidents on those days.

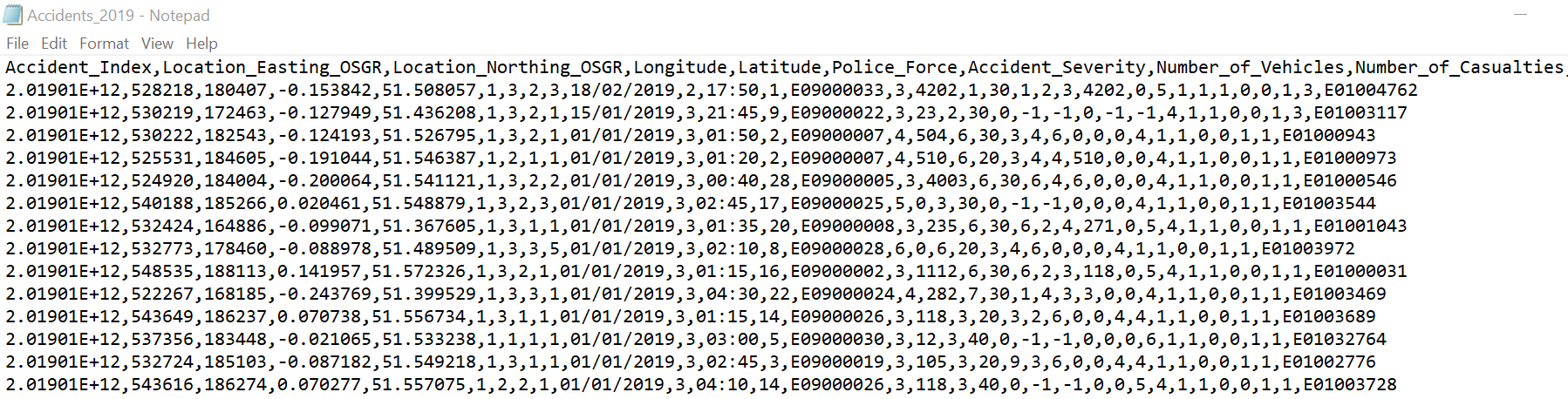
The data used in this research was taken from UK data repository that is responsible for recording datasets for various sectors of the UK government. The dataset includes the road accidents data from 1971-2019 (data.gov.uk, 2021).

# Methodology

## Dataset

To build a system that predict the likelihood of the road accidents in specific area in Great Britain at a certain time of the day we needed to analyze and investigate the historical accidents dataset. The DFT provide the road safety dataset which contain information and records about all the accidents that happened in Great Britain in the period from 1979 to 2019. We chose to work with the most recent data for the last 20 years from 1999 to 2019. The recent data have fewer missing data and can reflect the present better than the oldest data because it has less mutual factors with the recent road safety measures and this does not reflect the development in the vehicles manufacturing among other reasons. The original dataset consists of 8 csv files as follow (Figure 1 shows sample of dataset):

* Accidents\_1979\_2004
* Accidents\_2005\_2014
* Accidents\_2015, Accidents\_2016, Accidents\_2017, Accidents\_2018, Accidents\_2019
* Variable’s lookups.



*Fig 2. Initial dataset*

To extract the last 20 years dataset from the original dataset these steps were as follow:

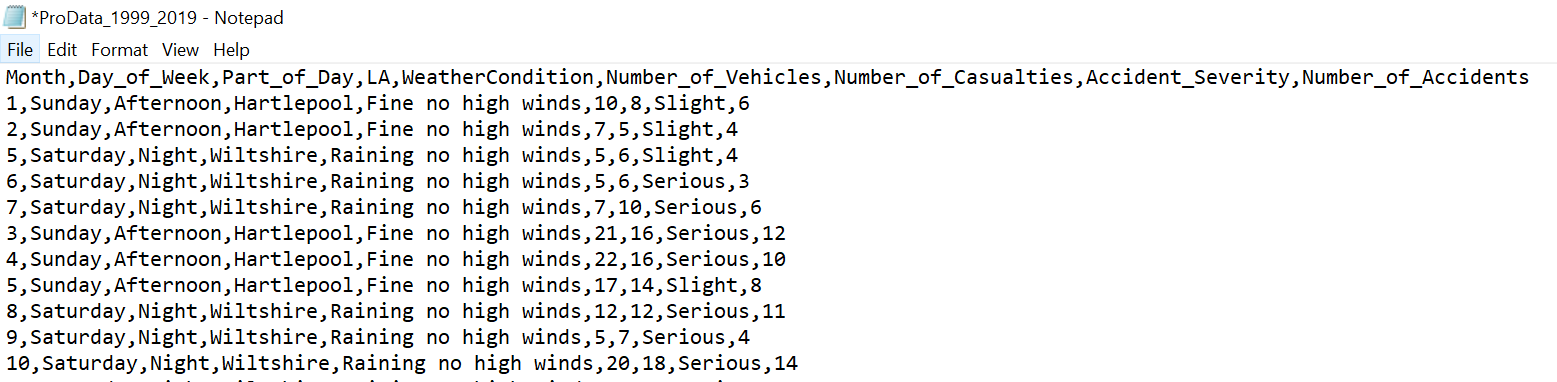
1. Split the first csv file to extract the data from 1999 to 2004.
2. Create a SQL server relational database using the dataset files and the variable lookups.
3. Upload all the accidents data from 1999 to 2019.
4. Create functions to read the variable names in the dataset.
5. Evaluate the dataset attributes according to three factors to manually select the attributes subset.

Table 1 shows the selection criteria and the purpose of each chosen attribute.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Contribution to the research analysis | Contribution to the chosen research level | Contribution to the prediction system |
| Accident\_Index | No | No | - |
| Location\_Easting\_OSGR | Yes | No | - |
| Location\_Northing\_OSGR | Yes | No | - |
| Longitude | Yes | No | - |
| Latitude | Yes | No | - |
| Police\_Force | No | No | - |
| Accident\_Severity | yes | yes | To extract the Output class |
| Number\_of\_Vehicles | yes | yes | To extract the Output class |
| Number\_of\_Casualties | yes | yes | To extract the Output class |
| Date | yes | yes | Input |
| Day\_of\_Week | yes | yes | Input |
| Time | yes | yes | Input |
| Local\_Authority\_(District) | yes | No | - |
| LA | yes | yes | Input |
| 1st\_Road\_Class | yes | No | - |
| 1st\_Road\_Number | yes | No | - |
| Road\_Type | yes | No | - |
| Speed\_limit | No | No | - |
| Junction\_Detail | No | No | - |
| Junction\_Control | No | No | - |
| 2nd\_Road\_Class | yes | No | - |
| 2nd\_Road\_Number | yes | No | - |
| Pedestrian\_Crossing-Human\_Control | No | No | - |
| Pedestrian\_Crossing-Physical\_Facilities | yes | No | - |
| Light\_Conditions | yes | No | - |
| Weather\_Conditions | yes | yes | Input |
| Road\_Surface\_Conditions | yes | No | - |
| Special\_Conditions\_at\_Site | yes | No | - |
| Carriageway\_Hazards | yes | No | - |
| Urban\_or\_Rural\_Area | yes | No | - |
| Did\_Police\_Officer\_Attend\_Scene\_of\_Accident | No | No | - |
| LSOA\_of\_Accident\_Location | yes | No | - |

1. Deal with the null values and substitute them with the mean or mode values where relevant.
2. Perform aggregate query against the dataset to group the accidents record by the input features (local authority where the accident happened, then by time, day of week, month, and weather condition.) as table 1 shows.
3. For each group, the query calculated the summation of the number of casualties and vehicles included in the accidents in addition to the number of accidents and the accidents severity to use it as intermediate input features to extract to output class.

Figure 2 shows the final dataset for the we will work on for the next steps. The dataset consists of 266K instances and summarize the accidents records for the last 20 years in GB. The GitHub page provided in the last part of this section includes the code and the detailed steps to reproduce these tasks.

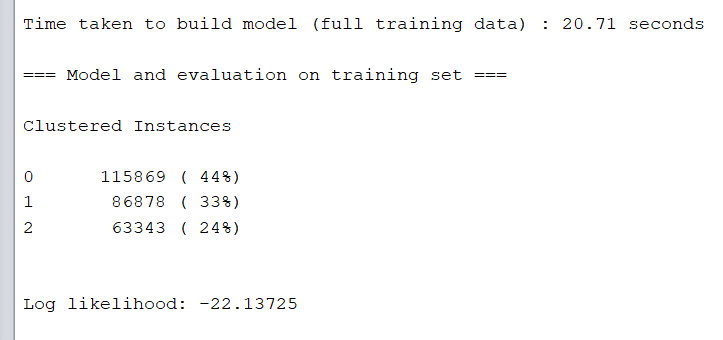


*Fig 2. Final dataset*

## 

## Output Class Analysis

The main function for our target system is to predict the likelihood of accidents in certain area with certain time and weather conditions. We run number of experiments on the final dataset using clustering machine learning algorithms to come up with natural clusters that represent the final class labels. Figure 3 shows one of the experiments results that classified the dataset instances to three clusters. For the next analysis step, we renamed these clusters (cluster0,1and 2) to be (Low, Medium, and High) to represent the accidents likelihood output class according to the intermediate input features values. For example, cluster0 (Low likelihood) represents the instances group where the number of vehicles, casualties and accidents are too few while clusrer2 (High likelihood) represents the instances with high number of vehicles, casualties, and accidents.



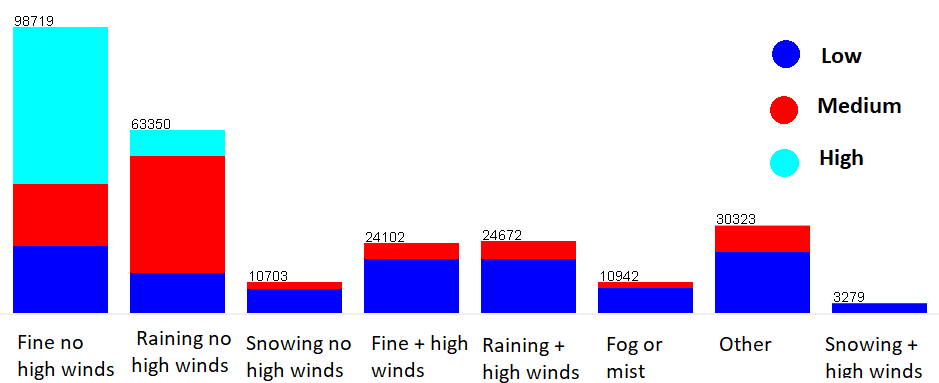
*Fig 3. Clustering algorithm*

The final dataset that would be used for the targeted prediction system consists of the input features and the output class after removing the intermediate features that were used in the clustering experiments.

# Module Experiments and Data Exploration

The target system takes the area name, time and weather conditions as input and predict the accidents likelihood in that area at three categorical levels: low, medium, and high likelihood. In this section, two machine learning classifications algorithms are introduced to build experimental prediction module. The experiments took place after data exploration and analysis to wisely select the suitable machine learning algorithms.

The dataset’s attributes are categorical that means it is suitable for the prediction analysis in general. The month input feature contains a seasonal information about the number of accidents and the accidents likelihood, but it does not show high correlation with the output class itself. On another hand, the weather conditions attribute showed very high correlation with output class where some values could be used directly to predict the output class. Figure 4 shows the weather conditions verse the accidents likelihood class. Six weather conditions values down grade the probabilities to two possible classes (low and medium) where the low represents the majority percent.



*Fig 4.* Classification by Weather

J48 and NaiveBayes classification Algorithms were used to build a prediction module. 70% of Dataset was used as training set while the modules were tested on the remaining data. The full experiments steps are described in the project GitHub page.

# Initial prototype / results (Results and Discussions)

Table 2 discusses the different Clustering and Classification algorithms that was used in this experiment.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | Clustered Instances | Observation | Results |
| EM | It provides 44% instances for the first cluster i.e., cluster 0, 33% for cluster 1, and 24% for cluster 2. | It was observed that whether the attribute number of accidents was normalised or not had negligible effect on the results. | In performing the algorithm, we noticed that when algorithm is run with number of vehicles as base attribute cluster 2, produces the greatest number of accidents, the results remain the same for number of accidents on roads. |
| SimpleKMeans | It provides 28% Instances for cluster 0, 38% for cluster 1, and 43% for cluster 2. | It was observed that this algorithm did not show many differences in the number of accidents caused in the different clusters, whether the attribute was normalised. | The cluster that recorded the highest number of accidents was cluster 1, with 820 accidents being the highest figure in the cluster, and slight differences occurred in results when the base attribute was changed. |

Table 3 discusses the Predictive algorithms that were used in this experiment.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Precision | Observation | Results |
| J48 | This algorithm correctly predicted 72.3865% instances from the provided dataset. | It is observed that prediction is made based on weather factors, and number of previously recorded accidents during those weather periods. | It classified the results into low, medium, and high were it made a prediction for the likelihood of accidents occurring in specified cities in the dataset. An interesting prediction was made which showed that contrary to our expectations the likelihood of accidents during snowy and raining weather conditions was exceptionally low. |
| NaiveBayes | 66.6404% instances were correctly predicted which is a little bit lower than that of J48. | ----- | It also made differentiation on classes of low, medium, and high which speaks of the likelihood of accidents occurring during various weather conditions. |

# Future work/reflection (Conclusions)

This research was done to predict the likely occurrence of an accident in a particular area or location. At the end of this report a predication system using machine learning was made to deduce the effect of weather and vehicle in the cause of most accidents. The data used for this research was gotten from the DfT public website and with the use of a relational database and data mining tools we made a prediction that clustered accident likelihood into (cluster0,1and 2) to be (Low, Medium, and High) within specific local areas and cities.

In this phase of the research, a model was developed using the combined data set from DfT between 1999 – 2019. This involved the use of clustering and trees for machine learning with specific conditions and a combination of several algorithms using attributes like weather and vehicle as well as how they relate to accidents in different locations or areas. These attributes have been shown to be reliable for accident predictions and future trends. In this research the predictions will be made considering weather, vehicle, road, and human factors. The predictions will be made for specific locations. We reviewed the results and compared the tables and found the Weather was the best result overall when compared in table 2 & 3.

In the next stage, we will be adding more data selection features, as well as providing a real time update and location of road traffic accidents, prediction in the future for the public to be aware and cautious. We will create a dashboard to portray this information making it easier and more accessible for the public to read and interpret the information.

# Bibliography

Data.gov.uk. 2020. Road Safety Data - Data.Gov.Uk. [online] Available at: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data (Accessed 07 March 2021).

Assum, T. (1997). Attitudes and road accident risk. *Accident Analysis & Prevention*, *29*(2), 153-159.

Brodsky, H., & Hakkert, A. S. (1988). Risk of a road accident in rainy weather. *Accident Analysis & Prevention*, *20*(3), 161-176. [Accessed: 10 March 2021]

Bergel-Hayat, R., Debbarh, M., Antoniou, C., & Yannis, G. (2013). Explaining the road accident risk: weather effects. *Accident Analysis & Prevention*, *60*, 456-465. [Accessed: 10 March 2021]

Elvik, R. (2000). How much do road accidents cost the national economy? *Accident Analysis & Prevention*, *32*(6), 849-851. [Accessed: 10 March 2021]

Taylor, M. C., Lynam, D. A., & Baruya, A. (2000). *The effects of drivers' speed on the frequency of road accidents*. Crowthorne: Transport Research Laboratory. [Accessed: 10 March 2021]

Reynolds, D. J. (1956). The cost of road accidents. *Journal of the Royal Statistical Society. Series A (General)*, *119*(4), 393-408. [Accessed: 10 March 2021]

Hakim, S., Shefer, D., Hakkert, A. S., & Hocherman, I. (1991). A critical review of macro models for road accidents. *Accident Analysis & Prevention*, *23*(5), 379-400. [Accessed: 10 March 2021]

Sagberg, F. (1999). Road accidents caused by drivers falling asleep. *Accident Analysis & Prevention*, *31*(6), 639-649. [Accessed: 10 March 2021]

Brown, I.D., 1994. Driver fatigue. Hum. Factors 36 (2), 298 – 314. Crone, R. (2016). Big Data Veracity Assessment: Improving risk assessment by adding high veracity data. to existing contents insurance models. PhD. Delft University of Technology, pp. 22 [Accessed: 10 March 2021]

Royal Society for Prevention of Accidents (2020). ‘Road Safety, Advice and information’, Available at: https://www.rospa.com/road-safety [Accessed: 10 March 2020]

Department for transport UK, Available at: https://www.gov.uk/government/publications/departmentfor-transport-single-departmental-plan/ [Accessed: 12 March 2020]

# Peer evaluation

|  |  |
| --- | --- |
| **Name** | **Tasks** |
| Ali Mohammed – 20002497 | * Helped with Issues. * Worked on methodology. * Worked on data set and results. * Worked on requirements. * Worked on module experiments and data exploration |
| Eniga Ahiante - 20026427 | * Helped with Issues and Risks. * Setup Microsoft SQL Server for data selection. * Upload raw data for processing. * Implement database scripts for data selection. * Setup GitHub for the group project. * Worked on the Initial prototype/Results section. * Worked on the problem statement. |
| Ethan Daniel – 17022218 | * Worked on the Literature review |
| Funmilayo Celestina Ayeni -19022115 | * Worked on the Literature review * Worked on the methodology section. * Worked on the data source section. * Helped with the problem statement section. * Worked on the Initial prototype/Results section. |
| Saba Rasheed - 20029056 | * Worked on the Aim and Objectives |
| Saif Javed - 16007638 | * Worked on the Introduction. * Worked setting the file and titles. * Worked on the problem statement. * Assisted with proving information on screenshots on results. * Worked on the Aims and Objectives. * Worked on the risks. * Worked on Abstract. * Worked on the Initial prototype/Results section. |