Abstract

The government is committed to enhancing traffic safety. Even though British roads are among the safest in the world, 127,967 reported injuries—of which 1,560 were fatal—and 27,300 serious injuries—were reported in 2021 alone. Our goal is to reduce incidents and make everyone's access to our roadways safer because every casualty is one too many.

To increase road safety, it is crucial to understand the root causes and contributing elements of events on our roadways. Although the evidence we now have is useful, we also know that additional work has to be done in order to better understand which strategies are most successful in eradicating them.

In this study, we are going to implement classification machine learning algorithms in order to determine the most important features affecting traffic accident severity and we will also conduct a statistical analysis on the road traffic accidents data between 2016 and 2021.

Table of contents

List of figures

List of tables

1. Introduction, Rationale and Objectives

In this chapter, the traffic accidents phenomena that this study aims to address is introduced and put into context. It also covers an ethical analysis, the precise goals and success criteria, and the details needed to access the source code repository.

1.1. Introduction

Since they cause numerous casualties, injuries, and deaths each year in addition to substantial economic losses, traffic accidents are one of the major global issues. Road accidents can be brought on by a variety of circumstances. It could be possible to take actions to lessen the effects and their severity if these elements can be better understood and predicted. The increasing development of road traffic has led to a surge in recent years in traffic accidents, particularly severe vehicle collisions. In reality, there has been a lot of focus in recent years on identifying the variables that have a substantial impact on the seriousness of traffic accidents, and this issue has been studied using a variety of methodologies.

The environment (such as weather patterns and traffic signs), the type of vehicle and its safety, and the characteristics of other road users are all elements that are associated to traffic accidents. Additionally, some of these characteristics play a bigger role in determining accident severity than others. It follows that analysis of the components that determine accident severity will help to reveal more patterns and information that can be applied to the avoidance of accidents.

1.2. Rational and user needs

The government is committed to enhancing traffic safety. Even though British roads are among the safest in the world, 127,967 reported injuries—of which 1,560 were fatal—and 27,300 serious injuries—were reported in 2021 alone. Our goal is to reduce incidents and make everyone's access to our roadways safer because every casualty is one too many.

To increase road safety, it is crucial to understand the root causes and contributing elements of events on our roadways. Although the evidence we now have is useful, we also know that additional work has to be done in order to better understand which strategies are most successful in eradicating them.

1.3. Objectives

The scope of this study is to give a better understanding of road traffic accident and the features causing different types of severity. This work will help bring into light the most important features contributing in making a traffic accident fatal, severe of slight.

To achieve this goal, three main objectives are determined:

1) Conduct a statistical analysis (Exploratory Data Analysis, EDA) of the traffic accidents datasets between January 2016 and June 2021 in order to:
   a) Examine the statistical characteristics of the data; namely, distribution and quality.
   b) Determine the pre-processing necessary to implement the machine learning algorithms.
2) Compare the machine learning algorithms on the level of the strengths, the weaknesses, the associated challenges, and opportunities in order to classify traffic accident data. The machine learning models used in this project are:
   a) Decision tree
   b) Random forest
   c) K nearest neighbour (KNN)
   d) Ada Boost
   e) Naïve Bayes
3) Test each algorithm's ability to fit test data.

1.4. Success criteria

The success of this study will be graded by two main criteria:

1) The predictive accuracy achieved by the best machine learning model.
2) The level of importance the data analysis will lead us to.

1.5. Ethical, legal, and professional considerations

The datasets used in this study are published under open source on:
https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

As such, there are no ethical, legal, or professional restrictions on its use. This work contains no personal data of any sort.

1.6. Code availability

All code used in this work is available at the following link:

https://gitlab.uwe.ac.uk/d2-debbagh/21059970_debbagh_driss_mscproject

2. Related work

2.1. Road traffic accidents statistics and analysis

Injury from traffic is a problem for health worldwide. Road traffic fatalities increased gradually over time, from 1.15 million in 2000 to 1.35 million in 2018. These numbers decreased in recent years, which can be misleading in most cases, due to the extraordinary circumstances observed because of the coronavirus pandemic (Fang-RongChang, He-Lai Huang, David C.Schwebel, Alan H.S.Chan, Guo-Qing Hu. 2020). Between 2010 and 2019, the number of vehicle kilometres driven in world increased each year. However, because of the significant decline in 2020, traffic projections are now below 2010 levels. Because the overall decrease is completely attributable to the dip in traffic levels seen in the 2020 predictions, it would be inaccurate to assert that traffic has decreased during the past ten years (https://roadtraffic.dft.gov.uk/summary).

Road traffic accidents are the eighth leading cause of mortality worldwide, accounting for around 2.37% of the world's 56.9 million deaths. As a result, the UN published the Global Plan for the Decade of Action for Road Safety 2011–2020 in 2011, and in 2015, it added the prevention of traffic injuries as Target 3.6 of the Sustainable Development Goals (SDG). However, in various nations, it has been shown that the data differ from reality. Evidence suggests that nations with the least capacity to address the societal, economic, and health service difficulties faced are disproportionately burdened with the current and predicted global cost of road traffic injuries. Despite the fact that the evidence basis used to make these estimations is still somewhat shaky and the quality of official data has recently declined due to the generally underdeveloped data systems in low- and middle-income nations. (Shanthi Ameratunga, Martha Hijar, Robyn Norton 2006)

The economic consequences of road traffic injuries include costs of prolonged medical care, loss of family breadwinner and loss of income due to disability, which together often push families into poverty. This, affect the economic wellbeing of the countries as well. The estimated 518 billion US dollars (USD) in global losses due to Road traffic injury (RTI) cost governments between 1% and 3% of their GDP, which is more than the total amount of development aid that these nations receive annually. The rise of RTIs globally has a major influence on many nation’s economies, especially those of low- and middle-income nations

(LIC and MIC), which frequently struggle to meet other development goals. LIC and MIC have greater road traffic fatality rates of over 90% despite having only 48% of the world's registered automobiles. (Koustuv Dalal, Zhiquin Lin, Mervyn Gifford, and Leif Svanström 2013).

2.2. Relationship between road accidents incidence and population size

Despite the fact that traffic accidents are a worldwide problem, it is still unclear how the number of accidents in cities varies with the size or density of the local population. The most popular method for assessing many elements of cities is to utilise straightforward per capita measurements. These, however, make the implicit assumption that urban traits rise linearly with population. This assumption is not very accurate since it fails to take into account the fact that cities with various populations have organisations and dynamics that are inherently nonlinear. According to Bettencourt and al's research (2010), these nonlinearities appear as scaling laws that demonstrate how agglomeration is an emerging phenomenon in metropolitan settings. So that if P represents the population size of the urban areas under consideration and I is an indicator of some sort, then:

$$I(P) = \alpha\, P^{\beta}$$

where the scaling exponent β is, in general, different from 1, and α is a proportionality constant. (Bettencourt LMA, Lobo J, Strumsky D, West GB. 2010)

On the research led by C. Cabrera-Arnau, R. Prieto Curiel and S. R. Bishop, they concluded that there is a sub linear relationship between population size and the risk that a person will experience at least one accident in an urban location during a certain length of time. But as the population grows, the chance that a particular accident would be fatal declines sub linearly. Therefore, it may be inferred that although larger conurbations have a higher accident probability, accidents there are generally less harmful.

Since traffic congestion can increase driver's stress levels, which, when combined with other urban stressors, would result in more accidents in urban areas with larger populations, it is hypothesised that this behaviour is indirectly caused by traffic congestion, which scales super linearly with city population sizes. The data reported here have significant policy-making ramifications because they appear to indicate a greater risk of road accidents in larger cities,

which is essential given the growing urbanisation the world is witnessing. (C. Cabrera-Arnau, R. Prieto Curiel and S. R. Bishop 2020)

2.3. Causes and the measures taken to control traffic accidents

Traffic accidents can be caused by carelessness, dangerous driving and bad habits. In the following, we are going to site some of the behaviours that leads to traffic accidents.

Distraction is one of the most frequent reasons for traffic accidents, and it may happen to any road user, regardless of their mode of transportation. Mobile phones, in-car entertainment systems, kids, and so-called backseat drivers are a few distractions.It is illegal to use a cell phone or sat nav while driving; this rule also holds true while you're stopped at a stop sign or stuck in traffic. This law carries a £200 fine and 6 licence points as a penalty. You will also lose your licence if you passed your driving test within the last two years. Everyone who uses the road has a responsibility for both their own and other people's safety. Always being on guard and aware is essential.

Speeding continues to be a major factor in traffic collisions. According to the Royal Society for the Prevention of Accidents (RoSPA), excessive speed is a factor in about 11% of incidents that result in injuries and 24% of collisions that result in fatalities that are reported to the police. Speed-related collisions are fully preventable; in order to protect other people and themselves, drivers and cyclists must abide by the rules of the road. It is important to always keep in mind that speed limits specify the top speed, not the minimum speed, therefore modifying your speed to the condition of the road or the weather is quite reasonable. One of the most crucial pieces of advice for drivers and riders is to maintain a safe distance from the car in front, also known as the "two second rule," in addition to being aware of your speed. The probability of an accident can be greatly decreased by keeping a safe distance from other cars.

Tiredness contributes to one in five auto accidents. Driving while fatigued makes it difficult to maintain attention, which puts both the driver and other road users in danger. When driving while feeling sleepy, you should pull over and take a break. Just make sure you do it somewhere safe—not on the hard shoulder of a highway, for instance. The AA advises drinking coffee or similar caffeinated beverage to increase attentiveness. If at all possible, try to get some rest so that you may resume your journey feeling revitalised.

Driving while under the influence of alcohol or drugs is prohibited and should be avoided at all costs. The NHS estimates that 9,050 individuals were killed or wounded in 2016 because at least one driver had consumed too much alcohol to drive. The legal alcohol limit for drivers is 80 milligrams in England, Wales, and Northern Ireland. The amount of alcohol one would need to consume to be considered above the legal limit for driving, however, varies from person to person. Driving while intoxicated or with a particular amount of illegal drugs in your blood is prohibited as well. Unfortunately, driving while intoxicated continues to be a major contributing factor in traffic collisions.

2.4. Literature review

Many researches were conduct in order to explain and determine the factors influencing traffic accidents, and specially the severity of the casualties. These researches, specially the recent made use of advanced analytical software and machine learning in order to help them achieve their goal.

The traffic accident dataset utilised in Tibebe Beshah Tesema, Ajith Abraham and Dejene Ejigu's article was received from the Addis Abeba traffic office, and the authors employed a genetic algorithm to construct a symbolic fuzzy classifier. The classifier that uses symbols to choose features from the accident dataset. The outcome demonstrates that the created classifier is capable of differentiating and categorising different types of injuries, and that the features utilised to label the data can be easily extracted and investigated. (Tibebe Beshah Tesema, Ajith Abraham and Dejene Ejigu 2012)

A machine learning experimental study using data on road accidents acquired in Ethiopia was also proposed by Tibebe Beshah Tesema, Dejene Eijigu and Ajith Abraham. They created a predictive model that explored the problem of data quality and projected the effect of driving behaviours on prospective injury risks using the CART, Random Forest, MARS, and Tree Net algorithms. The models can pinpoint the human-related causes of the severity of the disaster. The combined methods utilised in this paper were successful in terms of their ability to make accurate predictions. (Tibebe Beshah Tesema, Dejene Eijigu and Ajith Abraham 2011)

In Girija Narasimhan, Ben George Ben Ephrem and al's artical, a system was created to use predictive analytics with cutting-edge machine learning models to forecast the amount of accidents in Oman going forward. To improve the accuracy of the predictions, the author used a decision tree- and multiplicative-based boosted tree regression model. According to this

paper, human factors account for around 91% of all accidents as the primary or major contributing factor, with non-human variables accounting for the remaining 9%. (Girija Narasimhan, Ben George Ben Ephrem et al. 2017)

In order to investigate the association between fatal rate and other features including a drunk driver, light condition, collision mode, weather condition, and road surface conditions, L Li, S Shrestha and G Hu analyses the FARS dataset using the apriori, naive bayes, and k-means clustering algorithms. The factors that are strongly linked to fatal accidents are outlined in this research effort. The outcome demonstrates that a high fatality rate is caused by human factors such as drunk driving. (L Li, S Shrestha and G Hu 2017)

In contrast, the research in R Nidhi and V Kanchana work, used Nave Bayes and Apriori algorithms to identify patterns in the traffic accident. In this study, the authors created a prediction model based on the association rule to forecast the accident types that usually occur on new roadways. According to the study's findings, the majority of accidents involve vehicles that are less than five years old, and rural areas have a high death rate. (R Nidhi and V Kanchana 2018)

On the other hand, In the study led by Bulbula Kumeda; Fengli Zhang; Fan Zhou, they used a variety of classification methods, and the Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network, Multilayer Perceptron, and Nave Bayes classification strategies all produced results with excellent classification accuracy. They finally concluded that Fuzzy-FARCHD shows the best accuracy compared to the other algorithms. However, this study worked on a single year (2016) and did not use any of the vehicle data, which contains the drivers informations. A more appropriate approach is using a longer range of time in order to increase data, which will lead to a better performing algorithm, and utilising the vehicles data, which gives an understanding of the drivers and the vehicle's data. (Bulbula Kumeda; Fengli Zhang; Fan Zhou; Sadiq Hussain; Ammar Almasri 2019)

In order to create prediction models to evaluate injury severity, Beshah and Hill used Naive Bayes, Decision Tree (J48), and K-Nearest Neighbors classifiers. These models were then used to examine and predict the contribution of road-related elements to the severity of traffic accidents. Additionally, they used the WEKA tool and the PART algorithm to express the information as rules, with an accuracy rate of 79.94%. (T. Beshah and S. Hill 2010)

Using two years worth of crash data gathered in New Mexico, Chen and al. investigated the patterns of driver injury severity in rollover crashes using SVM models. The outcomes demonstrated that the polynomial kernel outperformed the Gaussian RBF kernel and that the support vector machine (SVM) models produce realistic predictions. G. Chen, Z. Zhang, R. Qian, R. A. Tarefder and Z. Tian 2016)

In order to predict traffic crashes, Dong et al. used two modules: an unsupervised feature and a supervised fine-tuning module. The findings shown that the feature learning section categorises interaction data between the explanatory factors and the feature representations, thereby reducing the input's dimensionality and maintaining the original data. (C Dong, C Shao, J Li and Z Xiong, 2018)

The effectiveness of the four machine learning algorithms to create accurate and dependable classifiers was examined in the study carried out by Rabia Emhamed AlMamlook, Keneth Morgan Kwayu, and Maha Reda Alkasisbeh. This covers the AdaBoost, Naive Bayesian Classifier, Logistic Regression, Random Forest, and Logistic Regression algorithms. Based on the test findings, it appears that the Random Forest performed better than the other models based on the confusion matrix F1-Score. According to this study's findings, algorithms can accurately forecast accidents 75.50% of the time. (Rabia Emhamed AlMamlook; Keneth Morgan Kwayu; Maha Reda Alkasisbeh; 2019).

3. Methods

This chapter outlines the methodological approach used in this study. It summarises the dataset, the ML models that aretested, and the metrics used in their evaluation.

3.1. Technical design and approach

In this study, we imported the dataset from data.gov.uk, then we conducted an exploratory data analysis with recoding the data, then we chose the features to include in the machine learning models by investigating the correlations and cramer's V. Then, by the end we implemented the Machine learning algorithms and chose the best performing one.

3.2. Dataset

In this study, we used data on road accidents that happened in the United Kingdom in the past five years, which includes the range between 2016 and 2020, then merged it with the first semester of 2021, that were taken from data.gov.uk. There was three different datasets: Accidents, Vehicles, and casualties. The attributes of each data set and a detailed description of it are defined in the tables below.

Tableau 1The attributes of the accident dataset and their description

Attribute	Attribute Description
No of vehicles	The total number of vehicles which takes part in the accident
Time (24hr)	The exact time when the accident occurs
1st Road class	A road where the accident occurred (Motorway, non-motor way …)
1st Road Class & No	A road which has a zonal system (A, B, C, Unclassified…)
Road Surface	The surface condition of the road during the accident (Dry, wet/Damp...)
Light Condition	A light condition during the accident (Daylight: street lights present…)

Weather Condition	Condition of the weather during the accident (Fine without high winds...)
Police force	The police force county which attended or included on their jurisdiction
Longtitude, Latitude	The longtitude and latitude of the location of the accident
Accident severity	The severity of the accident (Fatal, serious, slight)
No of casualties	The total number of casualties caused by the accident
Day of week	The day of the week when the accident occurred
time	The time of day when the accident occurred
Local authority district	The local authority district that the accident refer to.
Local authority district ONS	The local authority district that the accident refer to in the national office of statistcs references
Local authority highway	The local authority district that the accident refer to in the highway.
Road type	The type of the road where the accident happened (roundabout, one way street …)
Speed limit	The speed limit of the road where the accident happened
Junction detail	The details of the junction near the accident
Junction control	The type of control on the junction (stop sign …)
2nd road class	A road which has a zonal system (A, B, C, Unclassified…)
2nd road number	The reference number of the 2nd road
Special conditions at site	The special conditions at site if any

Carriageway hazards	Hazardous things on the road (previous accident …)
Urban or rural area	Urban, rural or undefined area where the accident happened

Tableau 21The attributes of the vehicle dataset and their description

Attribute	**Attribute description**
Vehicle type	The type of the vehicle (car, pedal cycle …)
Vehicle manoeuvre	The situation the vehicle was in (parked, reversing …)
Vehicle direction from	The direction which the vehicle was coming from (north, east, …)
Vehicle direction to	The direction which the vehicle was heading to (north, east, …)
Vehicle location restricted way	Which restricted way the vehicle was on if that applies (Bus lane, bus way …)
Junction location	The location of the junction (leaving main road, leaving roundabout …)
Skidding and overturning	Labelling if the vehicle skidded or overturned if that applies
Hit object in carriageway	Which object did the car hit in the carriageway if that applies
Hit object off carriageway	Which object did the car hit off the carriageway if that applies
First point of impact	First point of impact on the car (front, back …)
Vehicle left hand drive	Whether or not the vehicle is left hand drive
Journey purpose of the driver	The purpose of the journey the driver took
Sex	Sex of the driver
Age band	Age band of the driver (5 years band)
Engine capacity	Engine capacity of the car

Propulsion code	Which propulsion the car ride on (petrol, electric …)
Age of the vehicle	The age of the vehicle
Driver home area type	The type of area the driver is living in

Tableau 3 The attributes of the casualty dataset and their description

Attribute	**Attribute description**
Sex of casualty	Sex of the casualty
Age	Age of the casualty
Age band	Age band of the casualty (five years band)
Casualty severity	The severity of the injury of the casualty (Fatal, serious, slight)
Pedestrian location	The location of the casualty if he/she is a pedestrian
Pedestrian movement	The movement of the casualty if he/she is a pedestrian
Car passenger	The location of the casualty if he/she is a car passenger
Bus or coach passenger	The situation of the casualty if he/she on bus or coach
Casualty type	The type of the casualty (pedestrian, cyclist …)
Casualty home area type	The type of area the driver is living in

Every road traffic collision involving personal injuries that is reported to the police is documented in an administrative system with a statistical reporting component. Using Stats19 variables, this statistical reporting is carried out uniformly across Great Britain. The police officer responding to the accident often records the data in the MG NSRF form or a locally designed form (which contains the essential Stats19 variables).

In an additional third of situations, where a police officer has not responded to the personal injury accident, it is reported by members of the public at a police station some time after the

accident. According to STATS20 instruction, STATS19 data is gathered to a predetermined standard before being transmitted using STATS21 formatting and rules. To ensure that the data gathered by the police is pertinent to new requirements for road safety and lessens the burden on the police, the STATS19 standard is continuously evaluated. Every five years, reviews are carried out. The suggestions from reviewers alter the specification by including or excluding fields, including or excluding categories in fields, or by altering the data. 2018 saw the completion of the most recent STATS19 evaluation.

The Stats19 variable's goal is to deliver in-depth details regarding the accident's specifics, its surrounding circumstances, the participating vehicles, and the casualties that resulted. The information is verified using a variety of procedures (conducted by police departments, regional transportation agencies, and national governments), and is subsequently sent to the Department of Transport for use in national statistics. The information is also provided to regional highway agencies, who are mandated by law to promote driving safety.

The Stats 19 form is a series of forms:

- An accident record form: A document detailing the incidental circumstances surrounding this accident. These details can include things like the type, number, and speed limit of the road; the lighting, weather, and road surface; the presence or absence of intersections and facilities for pedestrian crossing; and the date, time, and location (by grid reference) of the accident.
- A vehicle record: For each car involved in the collision, a separate form is filled out, detailing the vehicle's movements prior to and during the collision as well as some personal information about the driver (age, sex, destination, and home postcode), as well as whether or not the collision was a hit-and-run.
- A casualty record: once more, each victim of the accident is listed separately on a separate form. There are details about the location and movements of any pedestrian casualties, the age and sex of the casualty, and the severity of their injuries (fatal, serious, or minor). There are also details about the casualty's class, or whether they were a driver or rider, a vehicle passenger (including car and bus passengers recorded separately), or a pedestrian.
- A contributory factors form: Each accident is given one of these forms, which includes a grid of 76 potential contributing elements and a space for the police officer to write

up to six of them that they believe are pertinent to the accident. Each of these variables is connected to one of the participants (either to a vehicle or a casualty, in which case a vehicle or casualty record will also be present; or to a "uninjured pedestrian," in which case no more information will be available). Additionally, the police officer notes whether the component was'very likely' to have caused the collision or only had a 'possible' connection. A single road user may be affected by multiple factors, and multiple road users may be affected by the same factor.

The police officer fills out a larger administrative form for each accident, which includes the Stats19 form. Additional fields describing the collision are included in this longer form and will be taken into consideration by the police when deciding whether or not to charge any of the parties involved. Additionally, it includes a description of the accident's location as well as the police officer's account of the incident's circumstances.

The Stats19 form is currently either a series of paper forms that the police officer fills out and then has keyed in by the police force's back office staff, or it is a digital version of the form that the police officer fills out directly. All around Great Britain, police departments are beginning to use handheld or other digital equipment to collect this data.

3.3. Data quality

Users may discover that contemporary accidents have more accurate or consistent information due to the ongoing improvement of the data flow and validation. Helping cops gather data has received a lot of attention and investment.

In the past, all police personnel would have used handwritten paper forms to collect data, which were not subject to the same amount of scrutiny as modern digital records. Some police officers also have access to superior data systems and mobile phone applications, which evaluate data entry as the officers are entering it. For instance, instead of relying on a written text description of the accident based on surrounding sites of interest, authorities can use the GPS on a cell phone to acquire precise coordinate information.

Over time, The Department for Transport (Dft) and Local Processing Authorities (LPAs) have also improved their validation methods, making it possible for them to spot problems earlier and send them directly to the reporting officer for assessment.

Additionally, users should think about whether abrupt changes in trends are caused by actual observable fluctuations or by other variables that alter specifications on an annual basis. Sharp increases can be seen if a field or category is added to a newer specification as more forces switch to the newer specification. As a result, more observations would be made as opposed to actually happening.

3.4. Machine learning modelling

3.4.1. Decision trees (DT) and Random Forest (RF)

1. Decision trees:

In graph theory, a tree is an undirected, acyclic and connected graph. The set of nodes is divided into three categories:

- Root node (access to the tree is through this node)
- Internal nodes: nodes that have descendants (or children), which are in turn nodes
- Terminal nodes (or leaves): nodes that have no descendants.

Decision trees (DTs) are a class of trees used in data mining and business intelligence. They use a hierarchical representation of the data structure in the form of sequences of decisions (tests) to predict an outcome or class. Each individual (or observation), which is to be assigned to a class, is described by a set of variables that are tested in the nodes of the tree. Tests are performed in the internal nodes and decisions are made in the leaf nodes.

Once the tree is built, classifying a new candidate is done by descending the tree, from the root to one of the leaves (which encodes the decision or the class). At each level of the descent, we pass an intermediate node where a variable xi is tested to decide which path (or subtree) to choose to continue the descent.

Principle of the construction:

Initially, the points of the learning base are all placed in the root node. One of the variables describing the points is the class of the point; this variable is called the "target variable". The target variable can be categorical (classification problem) or real value (regression problem). Each node is split, giving rise to several descendant nodes. An element of the learning base located in a node will be found in only one of its descendants.

The tree is built by recursive partition of each node according to the value of the attribute tested at each iteration (top-down induction). The optimized criterion is the homogeneity of the descendants with respect to the target variable. The variable that is tested in a node will be the one that maximizes this homogeneity.

The process stops when the elements of a node have the same value for the target variable (homogeneity).

2. Bagging:

The term bagging comes from the concatenation of the terms bootstrap aggregating. The concept is quite general and could be applied to different types of models. When applied to trees, it consists in making several trees from different subsamples of the training set and aggregating the results of all the trees.

Bagging is much more stable and better in terms of prediction than an individual tree. It also has the advantage that for each tree, there is an unused portion of the sample that can be used to evaluate the performance of each individual tree

3. Boosting:

Like bagging, boosting builds multiple trees and aggregates the results. Unlike bagging, boosting proceeds sequentially by changing the weights of the observations at each iteration.

We proceed as follows:

a. We set equal sampling weights for each observation
b. A random sample of size c is selected using the weights calculated in 1 or in the previous iteration.
c. Create a tree and make a prediction for each observation
d. We modify the weights in order to give more weight to the misclassified observations
e. Repeat steps 2 to 4 B times
f. Aggregate the results. In some cases, the weight of the trees in the final result can be adjusted according to the performance of the model.

3.4.2. Random Forests

We push the perturbation a little further: at each node of the tree, we select the variable among a sample of available variables.

We proceed as follows:

1. We select (generally with discount) a sample of data of size c
2. We build a tree b as follows:
    a. We start, at the root, with all the observations of the sample selected in 1.
    b. Select d variables from the set of available variables.
    c. For each of the d variables selected in ii, we calculate the Gini index (or another criterion) for each of the possible divisions (or combinations of modalities).
    d. We choose the variable and the division that brings down the Gini index the most and we create two new nodes on this basis.
    e. For each of the two nodes created, steps ii to iv are repeated until a certain stopping criterion is reached.
3. A prediction is computed for each observation
4. Repeat steps 1 to 3 B times (b=(1,2...,B))

Random forests are more robust than bagging because choosing a sample of variables at each node leads to a decorrelation of the trees. They are therefore generally more efficient while requiring less computation time than bagging.

However, the construction of a random forest implies the selection of one more hyperparameter: d, the number of trees selected at each node.

3.4.3. K nearest neighbour

One of the simplest supervised learning techniques for regression and classification is the k-nearest neighbour approach.

The model memorises the observations from the training set for the classification of the test set data using the "k-nearest neighbours" non-parametric method.

Because this algorithm doesn't pick up any new information during training, it is in fact considered lazy learning. It will search for its K nearest neighbours (using the Euclidean distance, or other methods) and select the class of the majority neighbours in order to predict the class of incoming input data.

The procedures to use this approach are as follows:

- The number of neighbours, k, is fixed.
- The k-neighbors that are most near the fresh input data that needs to be classified are found.
- By a majority vote, the matching classes are assigned.

3.4.4. Naïve Bayes

Naive Bayesian classification is a type of simple probabilistic Bayesian classification based on Bayes' theorem with a strong independence (called naive) of assumptions. It uses a naive Bayesian classifier belonging to the family of linear classifiers.

A more appropriate term for the underlying probabilistic model might be "statistically independent feature model".

In simple terms, a naive Bayesian classifier assumes that the existence of a feature for a class is independent of the existence of other features.

Depending on the nature of each probabilistic model, naive Bayesian classifiers can be trained effectively in a supervised learning context. In many practical applications, parameter estimation for naive Bayesian models is based on maximum likelihood. In other words, it is possible to work with the naive Bayesian model without worrying about Bayesian probability or using Bayesian methods.

Despite their "naive" design model and extremely simplistic basic assumptions, naive Bayesian classifiers have proven to be more than sufficiently effective in many complex real-world situations.

The advantage of the naive Bayesian classifier is that it requires relatively little training data to estimate the parameters necessary for classification, i.e. means and variances of the different variables. Indeed, the assumption of independence of the variables allows us to be satisfied with the variance of each of them for each class, without having to calculate a covariance matrix.

3.4.5. Ada Boost

The ensemble approach AdaBoost trains and deploys trees one after the other. By connecting a sequence of weak classifiers in AdaBoost, boosting is implemented. Each weak classifier attempts to correct data that were incorrectly classified by the weak classifier before it. Boosting accomplishes this by stringing together weak classifiers to produce a strong classifier. Because each decision tree has a tendency to be a shallow model that does not overfit but can be biassed, decision trees employed in boosting methods are known as "stump." A particular tree is trained to focus specifically on the shortcomings of the only other tree.

The weight of a sample that the prior tree incorrectly classified will be increased so that the following tree concentrates on correctly classifying the earlier incorrectly classified sample. When additional weak classifiers are sequentially added to the model, classification accuracy improves; however, this may result in severe overfitting and a loss of generalisation ability. AdaBoost works poorly when there is noise present, but it is well suited for imbalanced datasets. It takes longer to train AdaBoost.

3.5. Model evaluation

3.5.1. Confusion matrix

Tableau 4 Confusion matrix

	0 (predicted)	**1 (predicted)**
0 (observed)	TN	FP
1 (observed)	FN	TP

TN for true negatives predicted by the model

FP for false positives predicted by the model

FN for false negatives predicted by the model

TP for true positives predicted by the model

3.5.2. Evaluation metrics

From the confusion matrix, we can calculate many evaluation metrics:

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

All those metrics give values between 0 and 1, the closer it is to 1 the better is the model.

Each one of those evaluation metrics could be better to use depending on the context, depending on the importance of FP or FN we can chose between the precision and recall. However the F1 score and accuracy can give in general a good idea on the robustness of the models. In this work, we calculated all of those metrics.

4. Results

4.1. EDA results

4.1.1 Handling missing values

The datasets we have on our hands are accident, vehicle and casualty, containing 5 979 73 rows, 1 180 002 rows and 834 890 rows respectively.

In order to investigate the data quality in the three datasets (accident, vehicle and casualty), the table bellow shows the columns containing nan values, -1 values which means unreported since some of the accidents are self reported, and the solution proposed to manage them.

Tableau 5 Columns containing nan or -1 values and solution proposed

Dataset	Column	Nan values	-1 values	Solution
Accident	location_easting_osgr	692	0	Drop the column
Accident	location_northing_osgr	692	0	Drop the column
Accident	longtitude	42491	0	Drop the column
Accident	latitude	42491	0	Drop the column
Accident	Speed limit	37	172	The Nan and -1 values are merged with the unknown category
Accident	status	597973	0	Drop the column
Accident	Local authority district	0	43147	Included to the unknown category
Accident	Road type	0	1	Included to the unknown category
Accident	Junction detail	0	7	Included to the unknown category
Accident	Junction control	0	268585	Drop the column
Accident	Second road class	0	212827	Drop the column
Accident	Second road number	0	250811	Drop the column
Accident	Pedestrian crossing human control	0	834	Included to the unknown category

Accident	Pedestrian crossing physical facilities	0	784	Included to the unknown category
Accident	Light conditions	0	16	Included to the unknown category
Accident	Weather conditions	0	34	Included to the unknown category
Accident	Road surface conditions	0	1924	Included to the unknown category
Accident	Special conditions at site	0	839	Included to the unknown category
Accident	Carriageway hazards	0	870	Included to the unknown category
Accident	Urban or rural area	0	42359	Included to the undefined category
Accident	Did police officer attend scene of accident	0	8	Included to the unknown category
Accident	Trunk road flag	0	94656	Drop the column
Accident	Lsoa of accident location	0	42358	Drop the column
Vehicle	Vehicle direction from	0	8533	Included to the unknown category
Vehicle	Vehicle direction to	0	10473	Included to the unknown category
Vehicle	Age of driver	0	150377	Age out of bound
Vehicle	Engine capacity cc	0	340917	Engine capacity out of bound
Vehicle	Age of vehicle	0	339042	Age of vehicle out of bound
Vehicle	Generic make model	0	995915	Drop the column
Casualty	Sex of casualty	0	1999	Included to the unknown category

Casualty	Age of casualty	0	14924	Age of casualty out of bound
Casualty	Age band of casualty	0	14924	Included to the unknown category
Casualty	Pedestrian location	0	16	Included to the unknown category
Casualty	Pedestrian movement	0	18	Included to the unknown category
Casualty	Car passenger	0	1968	Included to the unknown category
Casualty	Bus or coach passenger	0	234	Included to the unknown category
Casualty	Pedestrian road maintenance worker	0	478	Included to the unknown category
Casualty	Casualty type	0	19	Included to the unknown category
Casualty	Casualty home area type	0	103254	Included to the unknown category
Casualty	Casualty imd decile	0	103618	Included to the unknown category

4.1.2. Recoding categorical features

The dataset imported from the gov.uk website, had numbers in the categorical variables referencing to the categories. So, in order to recode our dataset we used the data guide available on the website: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

4.1.3. Descriptive statistics

As for the target variable, accident severity, the accidents with a slight severity represents 80.4 % of our accident data, the serious ones represents 18.25 %, and finally the fatal ones represents 1.35 %.

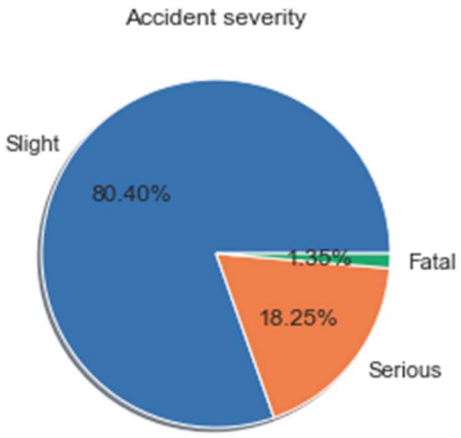


Figure 1 Accident severity distribution

The evolution of the accidents, as shown in the figure 2, shows a slight decreasing trend up until 2020 where the number of accidents decreased dramatically, then switched to an increasing trend.

On one hand, the mean of the number of accidents per month before 2020 is 10557.79, with a standard deviation of 898 and a minimum of 8838. On the other hand, after 2020, the mean drops to 7419.83, with a standard deviation of 1685.94 and a minimum of 3298, which is the lowest point in decades.

This trend is cause because of the COVID 19 pandemic, which led to a lockdown, as the investigation led by the government. This investigation found that, when compared to the 3-year average for 2017 to 2019, the number of traffic fatalities decreased by the greatest monthly percentage (68%) in April 2020. This is consistent with the drop in automobile traffic (63%) and the first full month of the nationwide lockdown. Compared to other categories of road users, the trend for deaths among pedal cyclists was different. However, the number of fatal pedal cyclist accidents climbed along with pedal cyclist traffic during the first lockdown.

This analysis revealed that the European Union exhibited a similar pattern, as evidenced by the 17% decline in road fatalities that Great Britain experienced in 2020 compared to 2019 and that was also seen across the European Union (EU) during the same time period. In 2020, Great Britain's decline followed a similar monthly pattern to that of other EU nations. The impact of travel restrictions may be seen in April, when there was the biggest percentage decline. (Adnan I.Qureshi, Wei Huang, Suleman Khan and Iryna Lobanova 2020)

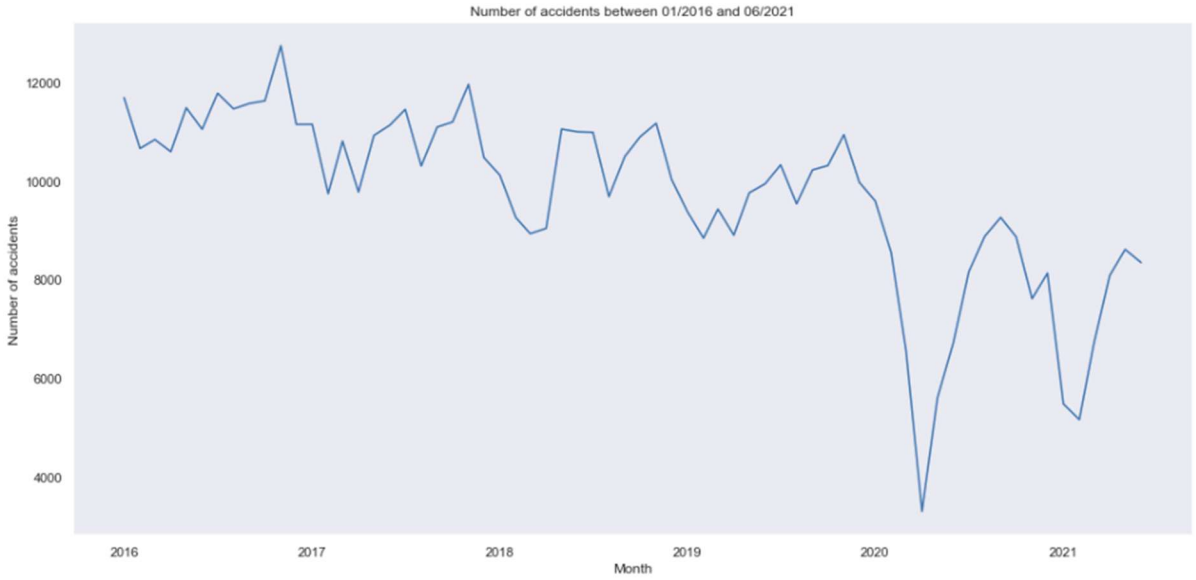


Figure 2 Evolution of the number of traffic accidents between January 2016 and June 2021

However, the number of fatal accidents did not follow the same trend as shown of the figure bellow. The trend decreased slightly, but did not hit the lowest point until 2021. On the research led by Adnan I. Qureshi, Wei Huang, and Suleman Khana (2020), they found that social lockdown policies do not reduce traffic accidents that result in catastrophic or fatal injuries, only those that result in non-serious or no injuries. Reduced commute and vehicle use during the required societal lockdown may be the cause of the decrease in traffic accidents that result in minimal or no injuries. Uncertainty surrounds why there was no decline in car accidents causing serious or fatal injuries under the ordered societal lockdown. Possible causes include a rise in traffic speed brought on by less congestion. Increased traffic speed could negate the benefits of less traffic by increasing the number of serious or deadly traffic accidents.

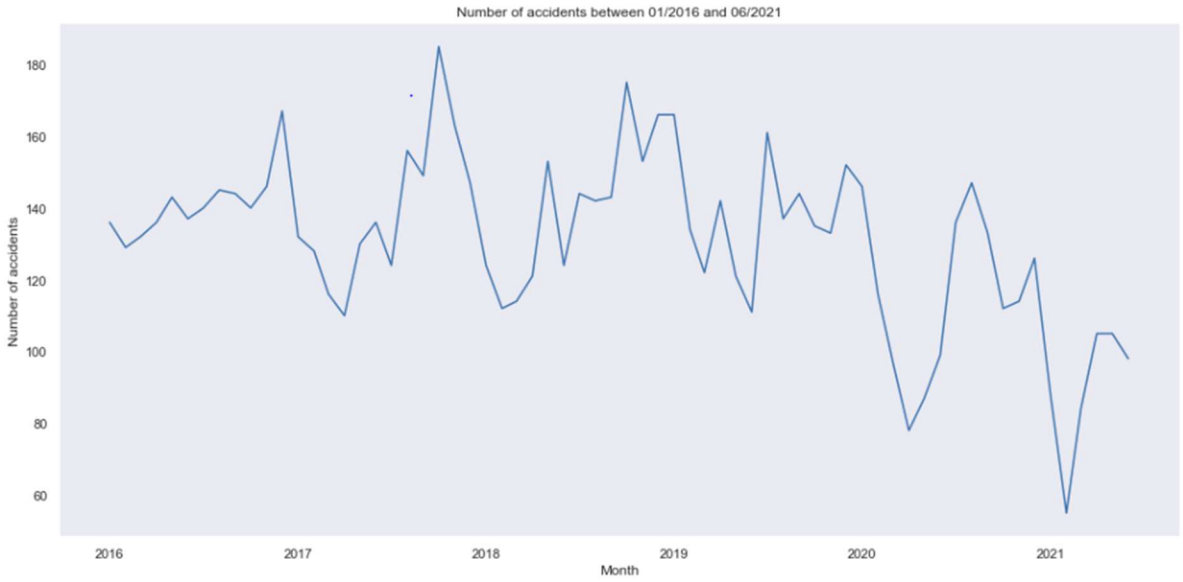


Figure 3 Evolution of the number of fatal traffic accidents between January 2016 and June 2021

When investigating the number of accidents occurring by day of the week, we find that, as shown on the table 5, the number of accidents increases during the working days of the week, hitting its maximum point on Friday, then decreases dramatically on the weekend and hits the lowest point on Sunday. Weekday accident rates are significantly higher than weekend accident rates. Due to the distinct differences between the geographical effects of factors like subways, schools, and hospitals on weekdays and weekends.

Tableau 6 Number, mean and standard deviation of the traffic accidents by day of the week

Day of week	No of accidents	Mean	Standard deviation
Monday	90256	44.95	113.73
Tuesday	94587	47.11	119.05
Wednesday	96163	47.89	121.04
Thursday	97615	48.61	122.77
Friday	104907	52.24	131.95
Saturday	84971	42.32	106.33
Sunday	71832	35.77	90.36

However, the number of fatal accidents does not follow the same trend as the maximum points meets in the weekend, and fluctuates over the week. Where the other two categories

follow the same trend. This is probably due to the increase of relentless driving during the weekend, which is amplified by the consumption of drugs, alcohol, and other substances.

Tableau 7 Accident severity percentage by day of the week

Accident severity	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Fatal	0.135	0.129	0.131	0.141	0.149	0.163	0.152
Serious	0.135	0.143	0.145	0.147	0.161	0.142	0.128
Slight	0.142	0.149	0.152	0.154	0.165	0.130	0.108

The distribution of the number of accidents by hour, as shown on the figure below, shows that there is two peak points, 8am and 5pm, which can described as rush hours, but still, the hours when the maximum of the traffic accidents occur are between 3pm and 6pm. This, in fact, fits exactly the working and the school hours, where the car traffic hits its highest point.

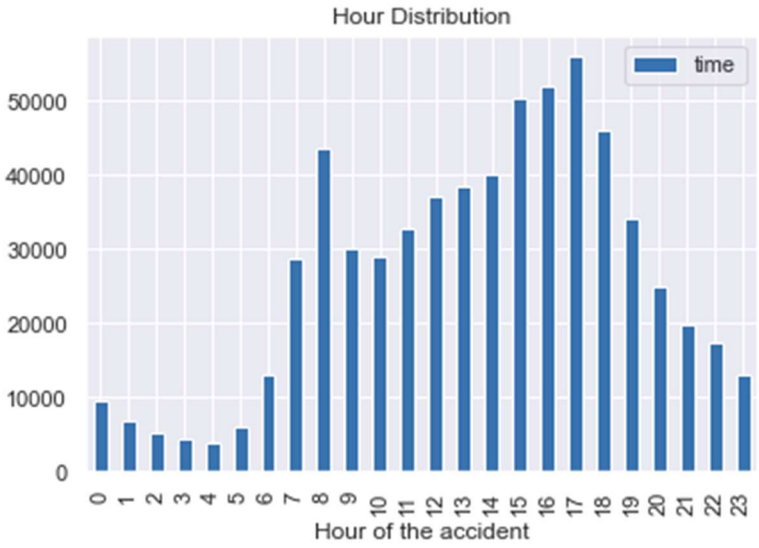


Figure 4 The distribution of the number of traffic accidents by the hour of the day

When we investigate the distribution of the number of accidents by hour on the weekdays, we find that during the working days, the distribution is slightly the same with the peak points at 8 am and 5 pm. However, on the weekends the distribution is different, where the peak points are between 12am and 6 pm.

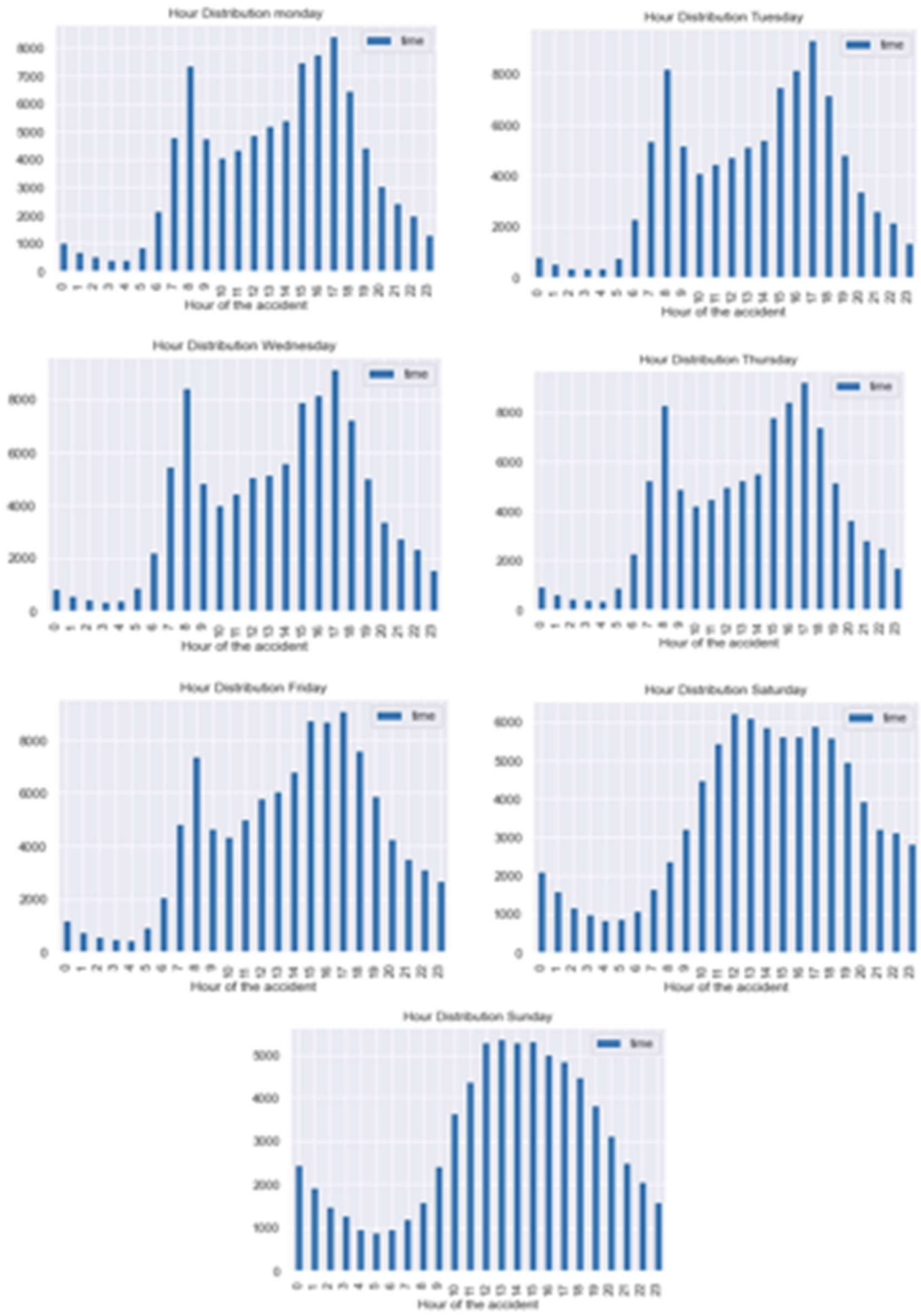


Figure 5 The distribution of the number of traffic accidents by the hour of the day for all days of the week

When investigating the hour distribution on the three categories of the accident severity (slight, serious and fatal), we find that their distribution is nearly the same. With a difference

on the fatal accidents, which does not have a peak point on 8am in contrary of serious and slight accidents, but they all have a peak point between 4pm and 5pm.

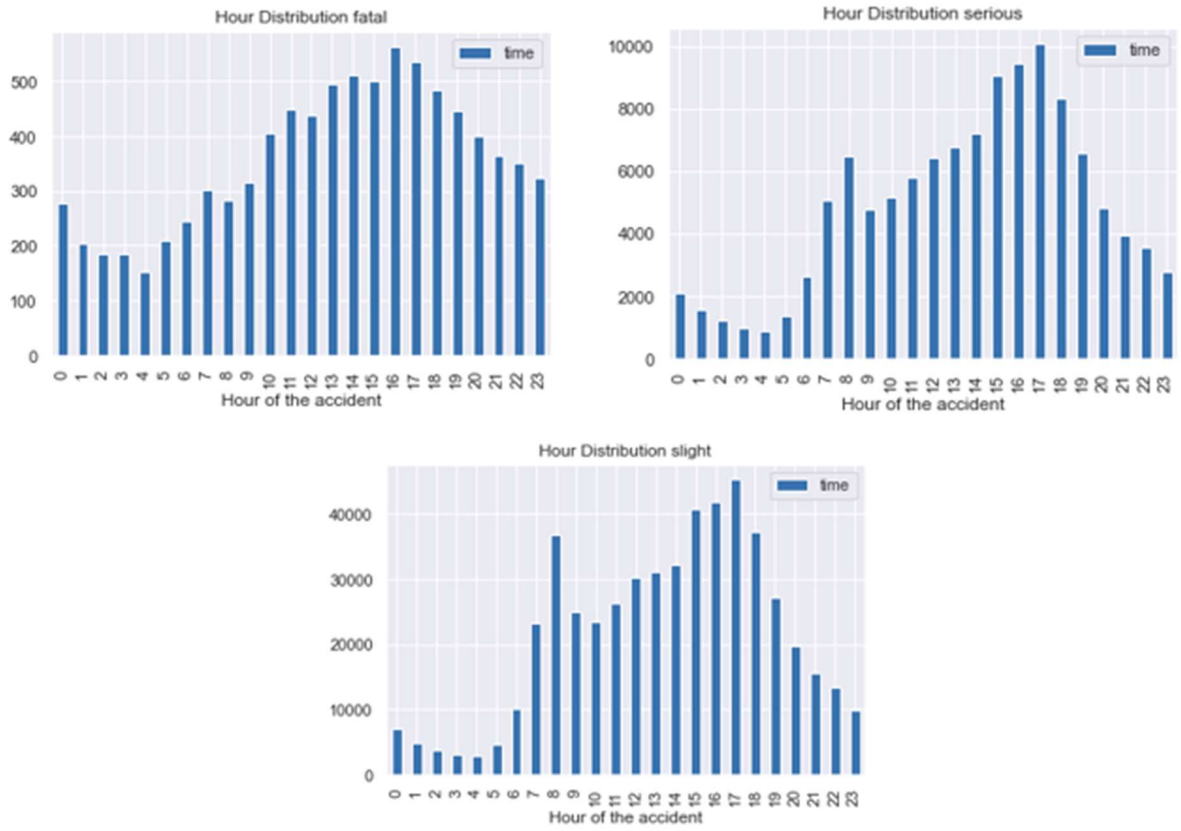


Figure 6 The distribution of the number of traffic accidents by the hour of the day for each severity type

The months with the most road accidents are January with 57389 accidents, and June with 58178 accidents, and with the least road accidents are April with 49678 road accidents and August with 49850. However, if we focus on the severity of the road accidents, we find that fatal accidents mostly occur on January and December (792 and 758 respectively), and the months with the least number of road accidents are February and March (674 and 665 respectively). For the serious accidents, they mostly occur on May and June with 10943 accidents and 11217 road accidents respectively, and the minimum points on February and December with 8907 accidents and 8727 road accidents respectively. Finally, for the slight road accidents, they mostly occur on January with 46874 road accidents, and June with 46256 road accidents, and with the minimum number of road accidents is seen on April with 39744 road accidents and August with 39526.

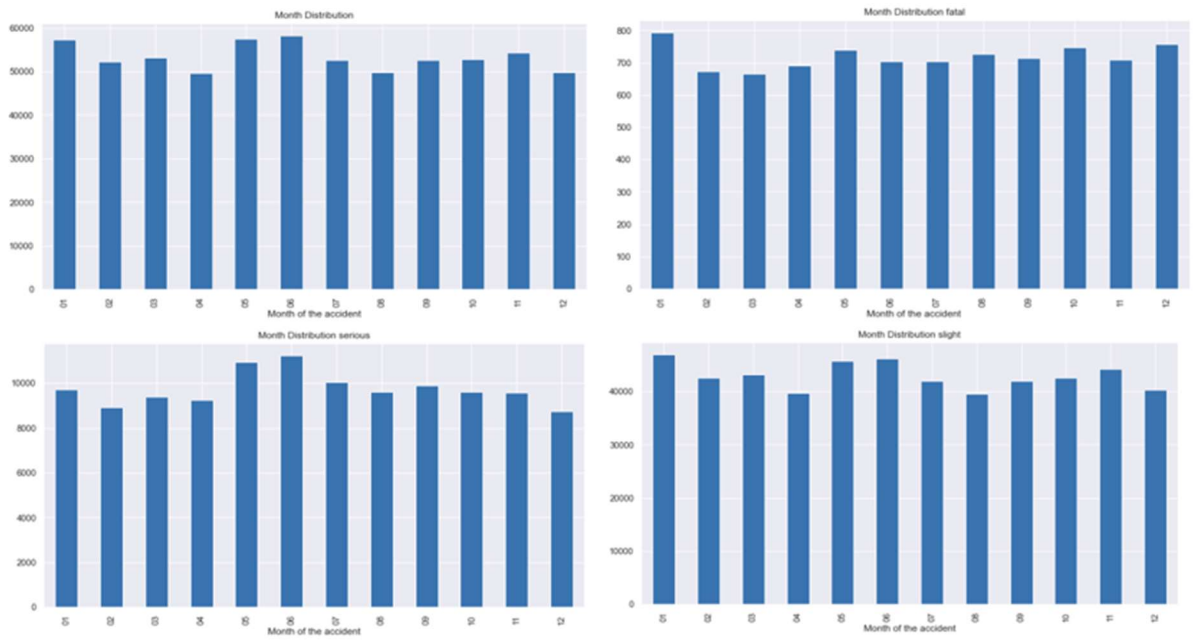


Figure 7The distribution of traffic accidents by month overall and for each severirty category

The number of vehicles on each road accident on the dataset on our hands varies between one and twenty-four, the majority of the road accidents happened between two vehicles 61.03 %, as shown on the table below, and the number of road accidents including one vehicle represents 29.01 % of our dataset.

Tableau 8The number and percentage of traffic accidents for number of vehicles taking part in the accident

Number of vehicles	Number of traffic accidents	Percentage
1	185794	29.01
2	390746	61.02
3	48892	7.63
4	10946	1.71
5	2590	0.40
6	788	0.12
7	333	0.05
8	121	0.02
9	56	0.0087
10	30	0.0046

11	12	0.0018
12	5	0.0008
13	5	0.0008
14	3	0.0005
15	2	0.0003
16	4	0.0006
17	1	0.0001
18	1	0.0001
23	1	0.0001
24	1	0.0001

The number of casualties on each road accident on the dataset on our hands varies between one and fifty-nine, the majority of the road accidents caused one casualty 79.47 %, as shown on the table below, and the number of road accidents causing two casualties represents 14.32 % of our dataset.

Tableau 9 The number and percentage of traffic accidents for each number of casualties

Number of casualties	Number of road accidents	Percentage
1	508858	79.4680
2	91689	14.3190
3	25722	4.0170
4	8967	1.4004
5	3160	0.4935
6	1150	0.1796
7	389	0.0607
8	163	0.0255
9	75	0.0117
10	61	0.0095
11	25	0.0039
12	22	0.0034

13	11	0.0017
14	6	0.0009
15	4	0.0006
16	3	0.0005
17	2	0.0003
18	1	0.0002
19	5	0.0008
20	2	0.0003
21	1	0.0002
23	2	0.0003
25	1	0.0002
26	1	0.0002
27	3	0.0005
29	1	0.0002
33	1	0.0002
34	1	0.0002
41	1	0.0002
42	1	0.0002
52	1	0.0002
58	1	0.0002
59	1	0.0002

The single carriageway carries most of the traffic accidents, and the dual carriageway comes on the second place. The same trend is observed when investigating the crosstab between the road type and the severity of the accident as shown on the figure 8 and the table X.

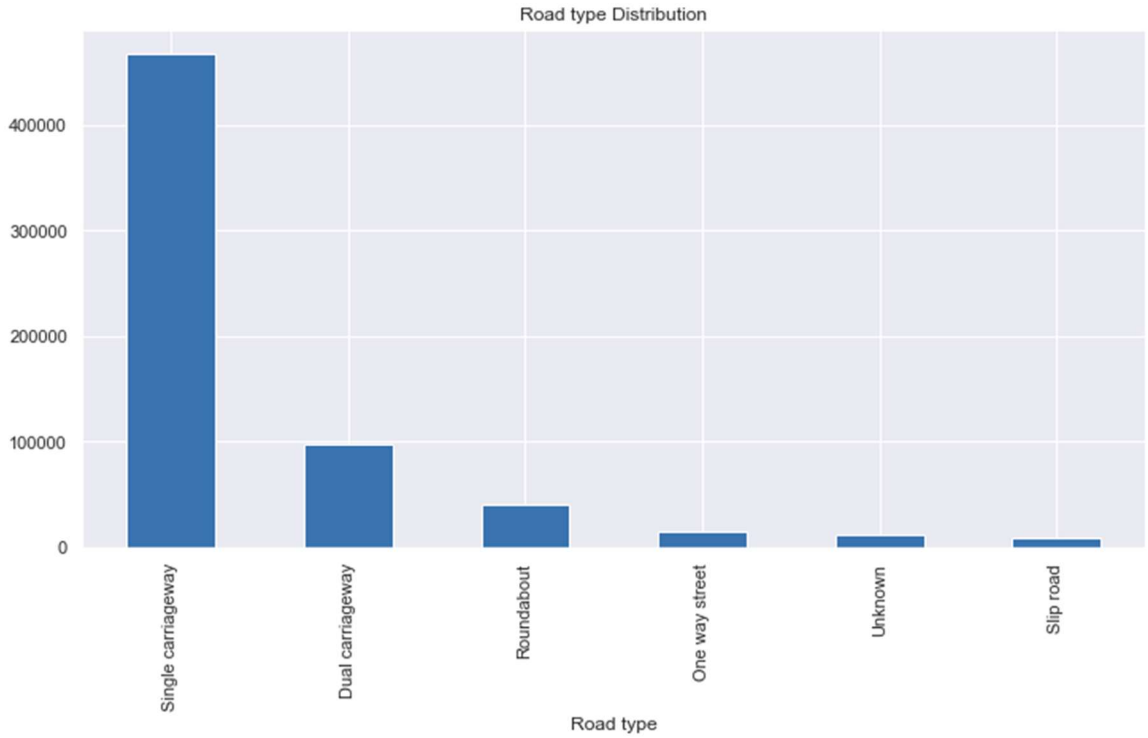


Figure 8 The distribution of traffic accident by road type

Tableau 10 crosstab between the road type and the severity of the accident

Accident severity	Roundabout	One way street	Dual carriageway	Single carriageway	Slip road	Unknown
Fatal	0.0166	0.0105	0.1992	0.7585	0.0104	0.0048
Serious	0.0451	0.0203	0.1375	0.7777	0.0098	0.0095
Slight	0.0668	0.0252	0.1559	0.7172	0.0144	0.0205

On our dataset, the urban areas caries most of the instances, since they have more vehicles and more traffic than the rural areas.

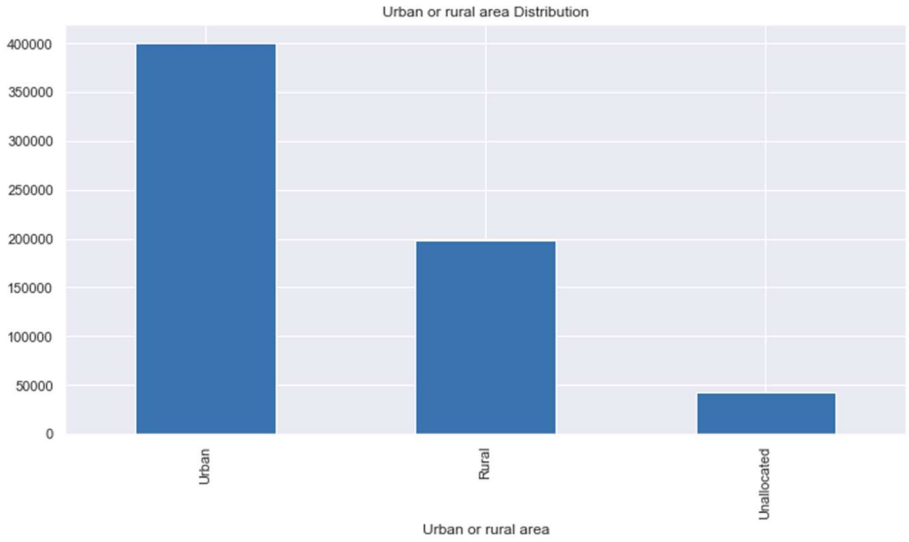


Figure 9 distribution of traffic accidents between urban and rural areas

However, most of the fatal accidents happen in rural areas 58.7 %, contrary to the serious and slight accidents, which mostly happen in urban areas, 55.8 % and 64.4 %. This is probably due to the fact that in and around areas the speed limit is high and the control measures are rare which leads to irresponsible driving and speed excess.

Tableau 11 Crosstab between accident severity and type of the area

Accident severity	Urban	Rural	Unallocated
Fatal	0.351	0.587	0.062
Serious	0.558	0.367	0.075
Slight	0.644	0.292	0.064

Most of traffic accidents happen on the roads with a speed limit of 30 miles per hour, which makes sense since most of the roads on the urban areas have the same speed limit. And the roads with the least number of traffic accidents are the ones with 50 miles per hour as speed limit.

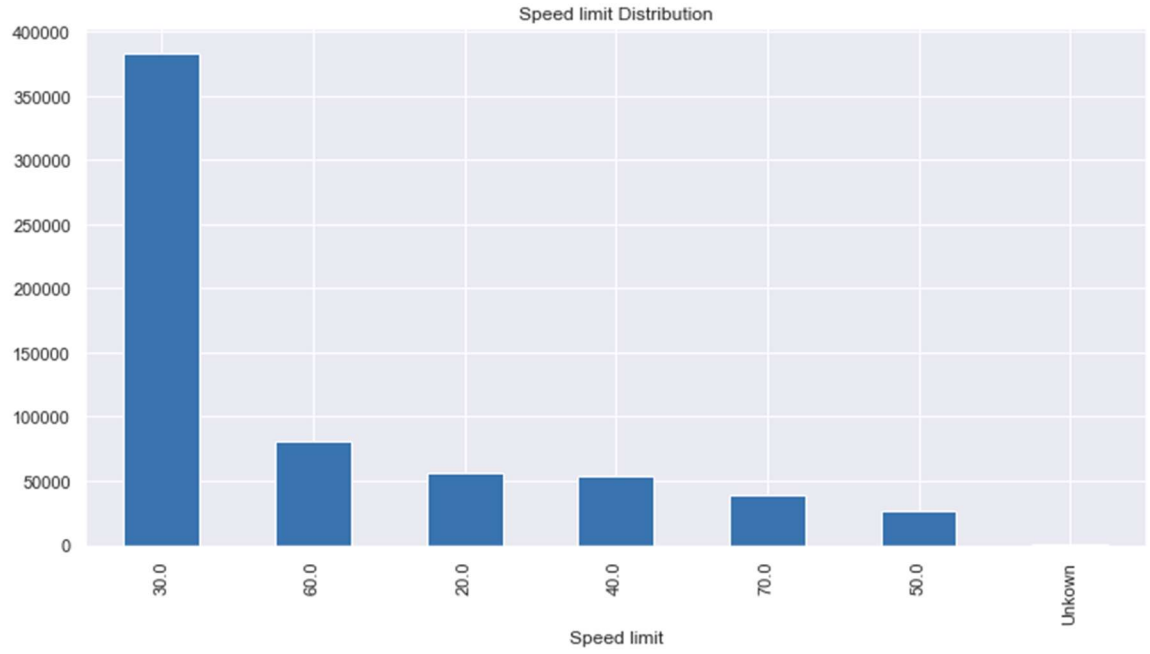


Figure 10 the distribution of traffic accidents by speed limit

However, most of the fatal traffic accidents happen on the roads with the speed limit of 30 and 60 miles per hour (33.76 % and 33.55 % respectively), when serious and slight ones mostly happen on the roads with a speed limit of 30 miles per hour.

Tableau 12 Percentage of the severity of traffic accidents by speed limit

Accident severity	20	30	40	50	60	70
Fatal	0.0296	0.3376	0.0983	0.0793	0.3355	0.1198
Serious	0.0749	0.5557	0.0894	0.0442	0.1778	0.0579
Slight	0.0918	0.6127	0.0826	0.0397	0.1114	0.0614

Most traffic accidents occur on daylight, with darkness and streetlights present and lit on the second position. This makes sense since the daylight time sees more traffic and most of the road in the United Kingdom that witnesses a lot of traffic has streetlights lit.

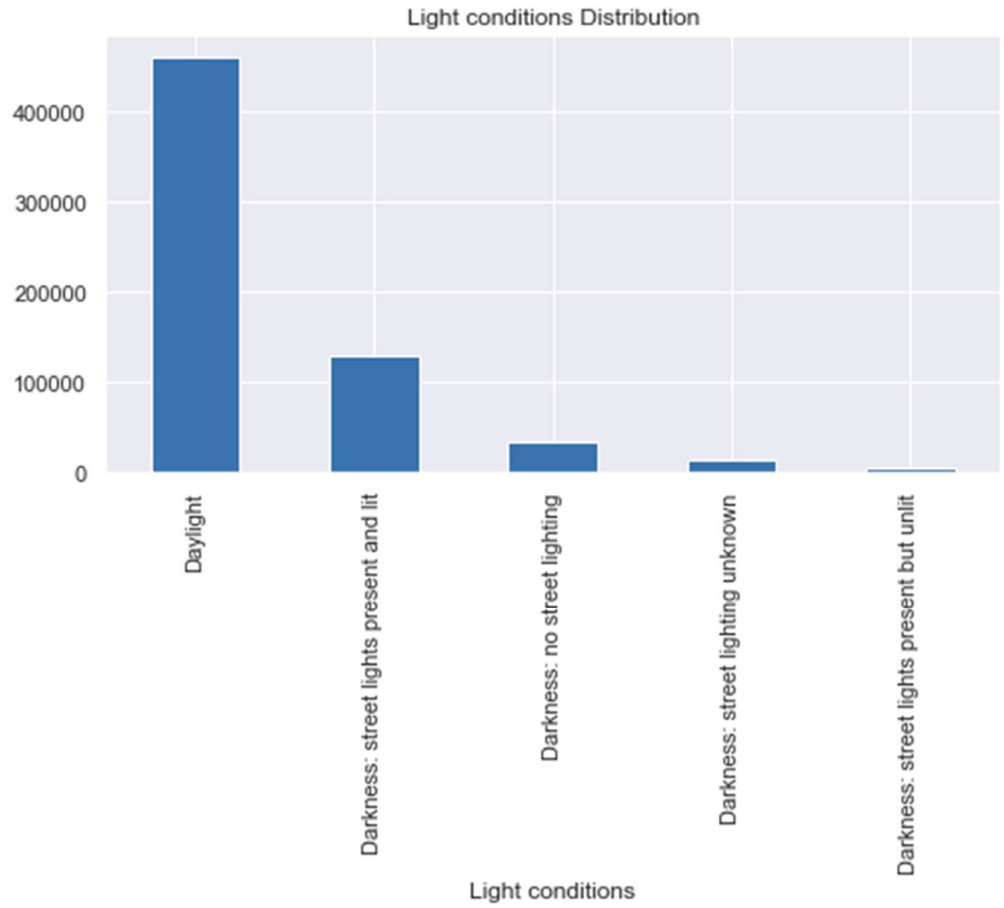


Figure 11The distribution of trafic accidents by light conditions

Tableau 13 Percentage of traffic accidents severity by light conditions

Accident severity	Daylight	Darkness: street lights present and lit	Darkness: street lights present but unlit	Darkness: no street lighting	Darkness: street lighting unknown
Fatal	0.589	0.202	0.012	0.179	0.018
Serious	0.696	0.211	0.008	0.070	0.016
Slight	0.728	0.199	0.007	0.044	0.022

The majority of traffic accidents happen on a fine without high winds weather, since most of the drivers take extra precaution when it is foggy or raining and drives slower with more attention taking into account the risk.

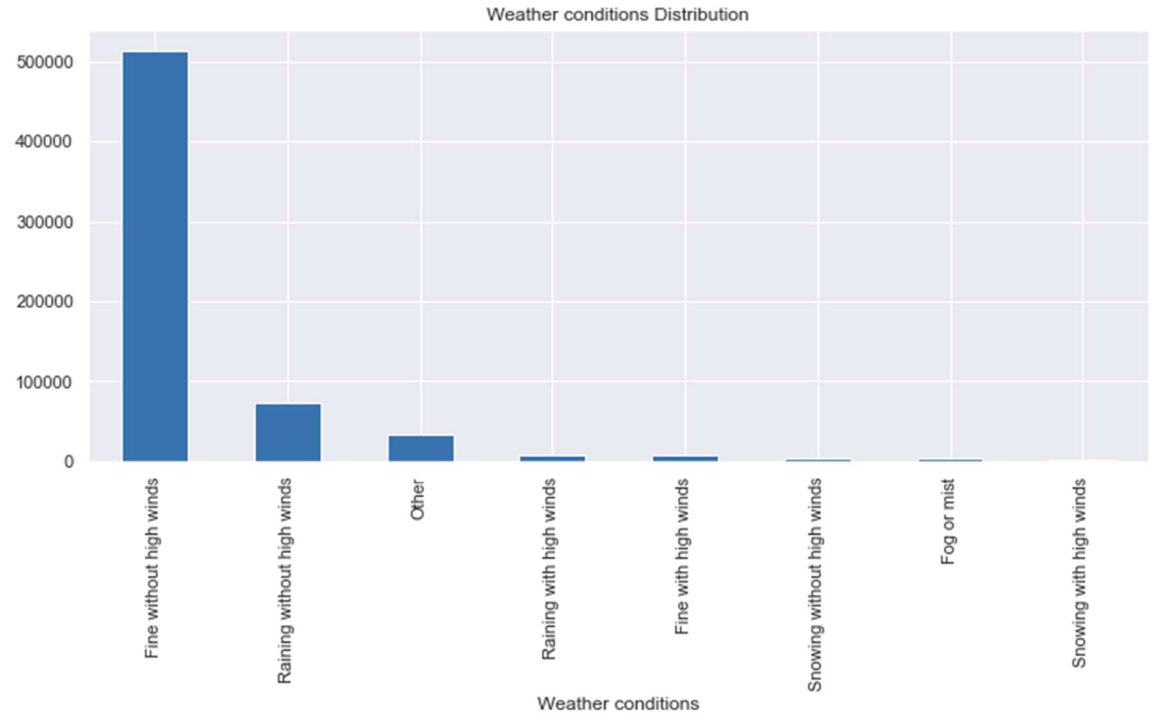


Figure 12 Distribution of traffic accidents by weather conditions

Tableau 14 Percentage of severity of road accidents by weather conditions

Accident severity	Fine without high winds	Raining without high winds	Snowing without high winds	Fine with high winds	Raining with high winds	Snowing with high winds	Fog or mist	Other
Fatal	0.8202	0.0985	0.0044	0.0168	0.0182	0.0003	0.0108	0.0307
Serious	0.8191	0.1077	0.0041	0.0119	0.0129	0.0010	0.0049	0.0385
Slight	0.7976	0.1149	0.0050	0.0101	0.0110	0.0012	0.0043	0.0560

Similarly to the weather conditions, drivers take extra care when the road is wet of it is snowing. That is why most of the traffic accidents occur on dry roads.

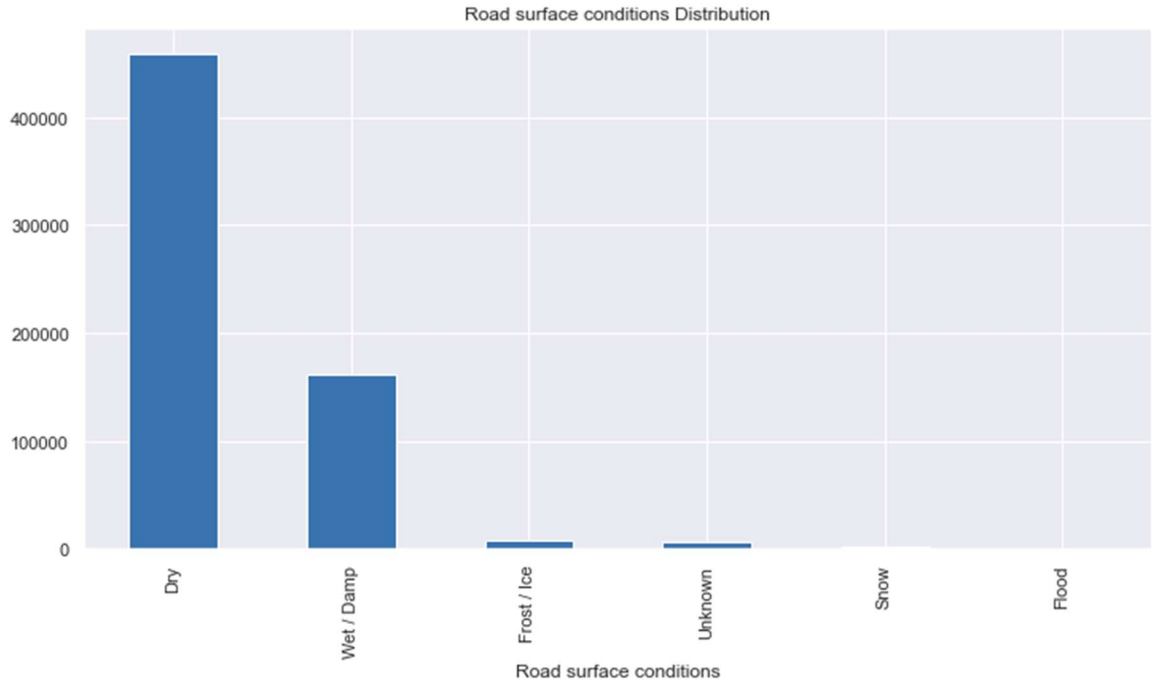


Figure 13 Distribution of traffic accidents by road surface conditions

Tableau 15Percentage of severity of traffic accidents by road surface conditions

Accident severity	Dry	Wet / Damp	Snow	Frost / Ice	Flood	Unknown
Fatal	0.6894	0.2945	0.0023	0.0108	0.0016	0.0014
Serious	0.7198	0.2584	0.0031	0.0125	0.0013	0.0049
Slight	0.7168	0.2515	0.0043	0.0137	0.0013	0.0125

The majority of the vehicles taking part in an accident on our data has men as drivers, this generalise also on all categories of the severity of the accident.

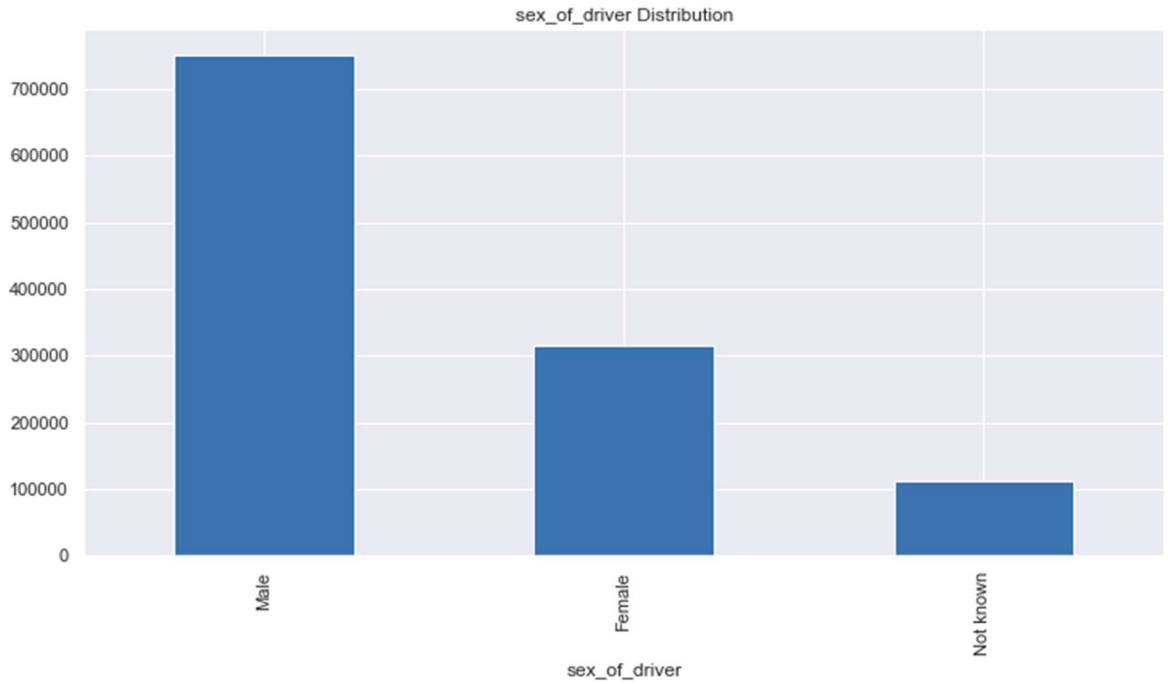


Figure 14 The Distribution of the vehicles by sex of the driver

Accident severity	Male	Female	Not known
Fatal	0.79	0.18	0.03
Serious	0.69	0.23	0.08
Slight	0.60	0.28	0.12

Figure 15 Percentage of accident severity by sex of the driver

The drivers aged between 26 and 35 are most likely to be involved in a road accident, and the band between 36 and 45 comes in the second place. This is probably because of lack of experience and or the consumption of drugs and alcohol while driving.

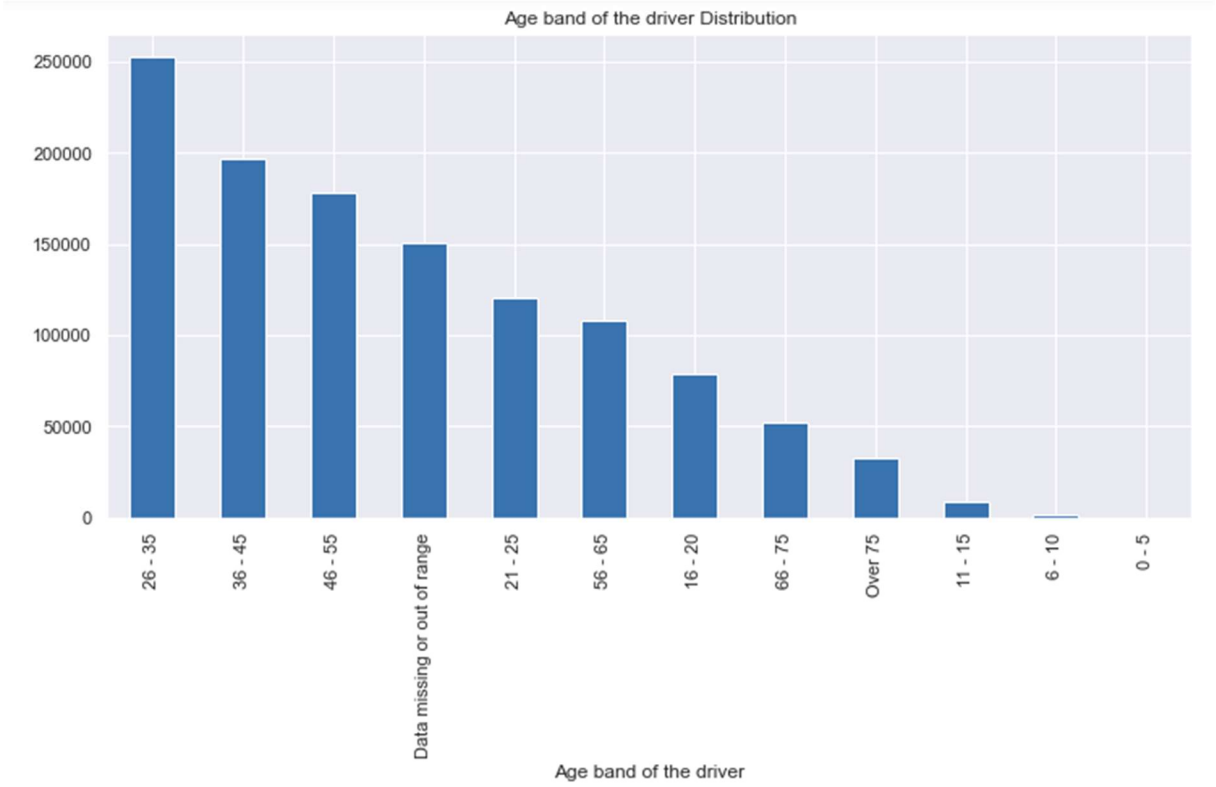


Figure 16 The distribution of vehicles by age band of the driver

However, we can see from the table 15 that, the drivers aged between 46 and 55 represents the majority of the drivers involved in a fatal traffic accident with 18.13 %. Moreover, the band age of drivers between 26 and 35 are involved in serious and slight traffic accidents.

Tableau 16 Percentage of accident severity by age band of the driver

Accident severity	0 - 5	6 - 10	11 - 15	16 - 20	21 - 25	26 - 35	36 - 45	46 - 55	56 - 65	66 - 75	Over 75
Fatal	0.0004	0.0000	0.0039	0.0576	0.0942	0.1794	0.1549	0.1813	0.1412	0.0767	0.0677
Serious	0.0002	0.0019	0.0088	0.0708	0.0964	0.2092	0.1586	0.1615	0.1072	0.0549	0.0363
Slight	0.0002	0.0018	0.0081	0.0611	0.0949	0.2176	0.1650	0.1451	0.0880	0.0403	0.0245

Most of the casualties on the dataset we are working on are male.

Tableau 17 Distribution of casualties by sex

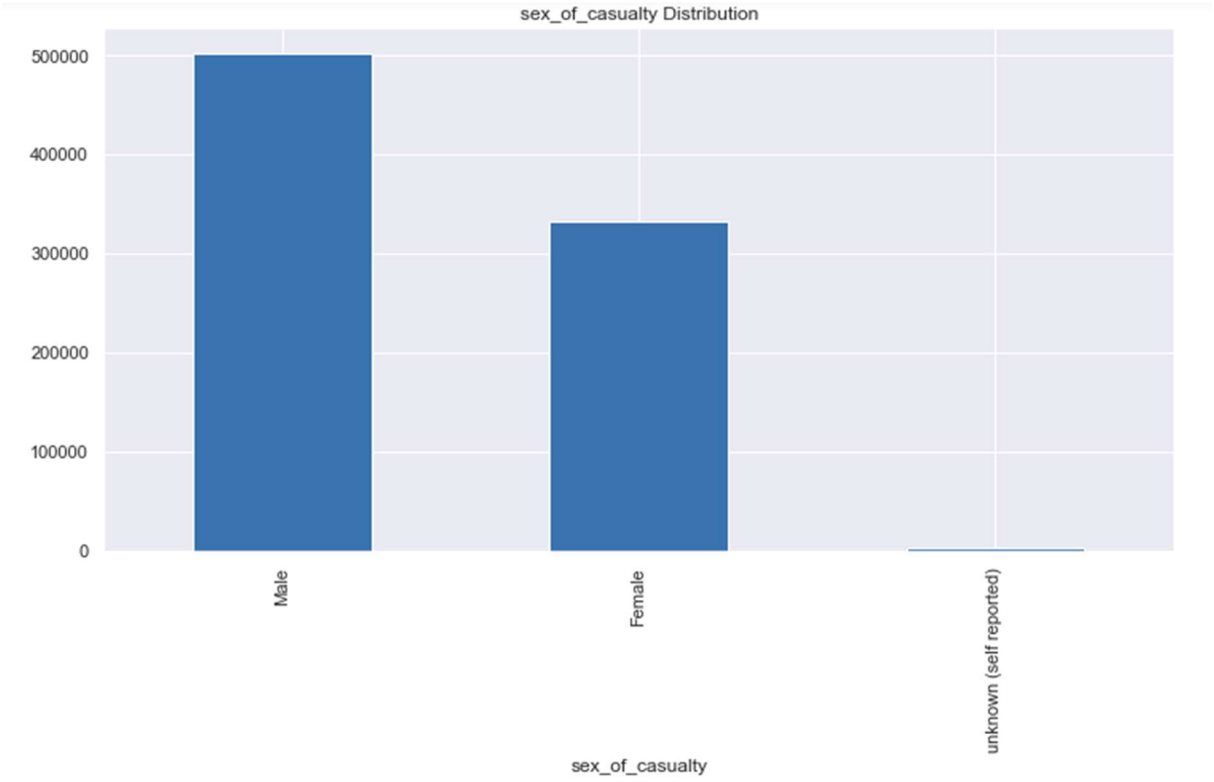


Tableau 18Percentage of casualty severity by sex

Casualty severity	Male	Female	unknown (self reported)
Fatal	0.7505	0.2495	0.0000
Serious	0.6931	0.3066	0.0003
Slight	0.5812	0.4159	0.0028

Most of casualties on our dataset are aged between 26 and 35. This distribution is also seen when crossing the casualty severity with the age band.

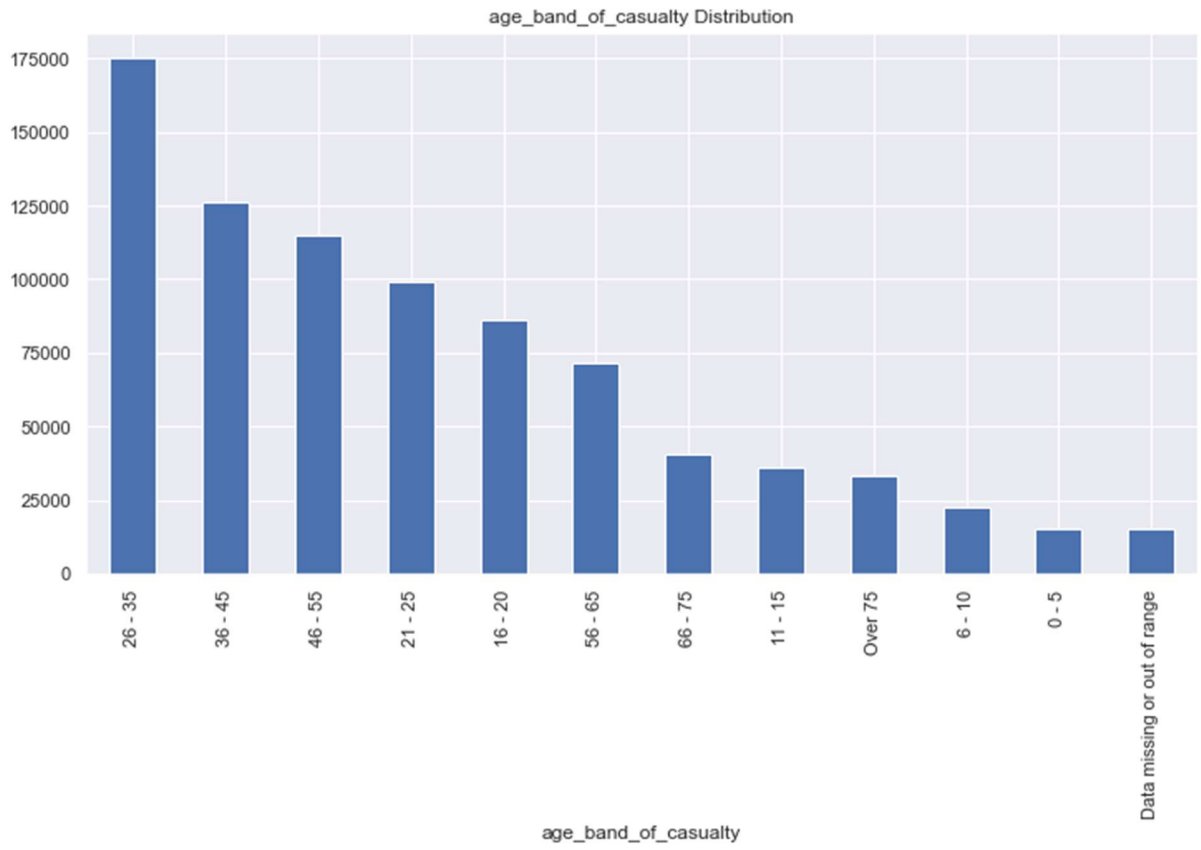


Figure 17 Distribution of casualties by age band

Tableau 19 Percentage of casualty severity by age band

casualty severity	0 - 5	06-10	11-15	16 - 20	21 - 25	26 - 35	36 - 45	46 - 55	56 - 65	66 - 75	Over 75
Fatal	0.01	0.01	0.01	0.08	0.10	0.17	0.12	0.13	0.12	0.10	0.15
Serious	0.01	0.02	0.05	0.11	0.11	0.18	0.13	0.14	0.10	0.07	0.06
Slight	0.02	0.03	0.04	0.10	0.12	0.22	0.16	0.14	0.08	0.04	0.03

The majority of the casualties are drivers or riders. This trend is also seen when crossing the the casualty severity by the casualty class. However, pedestrians shows a big percentage as serious casualties, which means that the authorities needs to take more measures in order to protect them.

Tableau 20 Distribution of casualties by casualty class

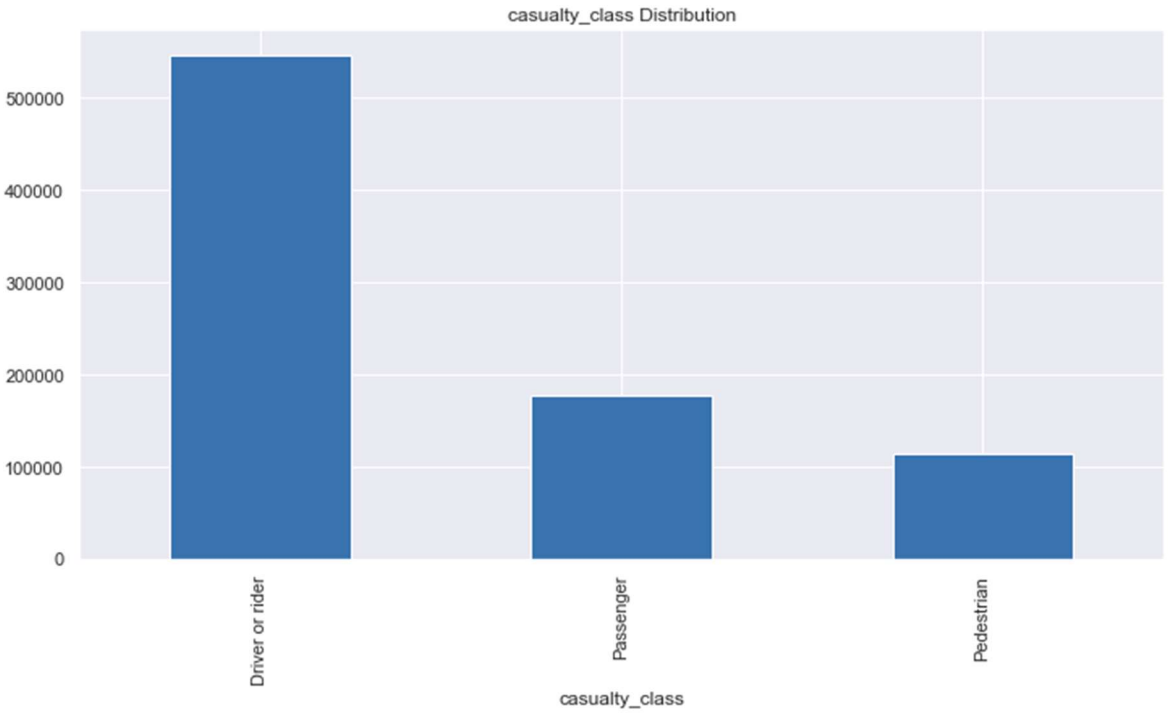


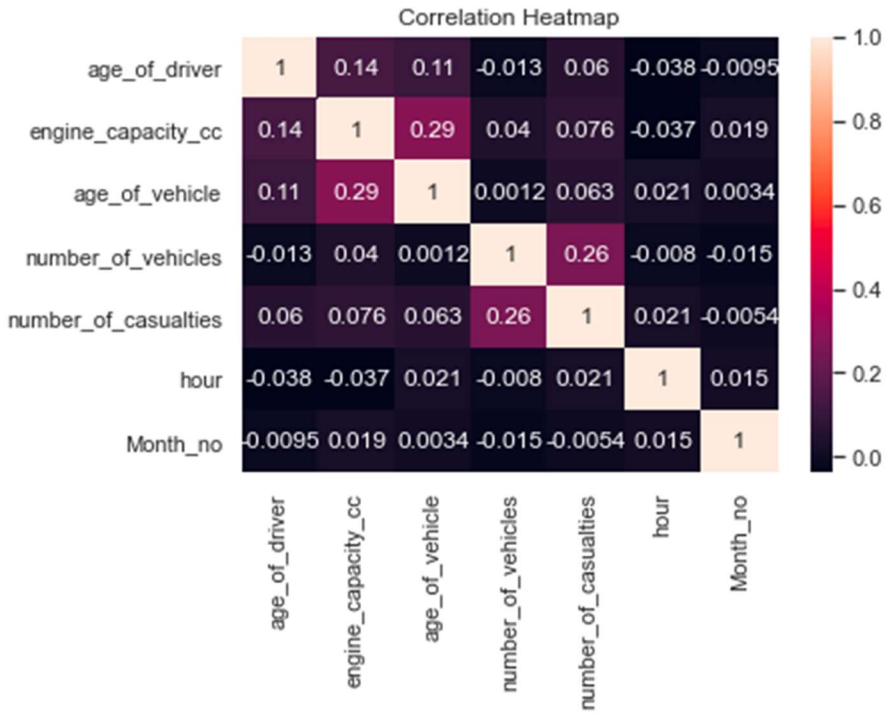
Tableau 21 Percentage of casualty severity by casualty class

casualty severity	Driver or rider	Passenger	Pedestrian
Fatal	0.60	0.15	0.25
Serious	0.64	0.14	0.22
Slight	0.66	0.22	0.12

4.1.4. Correlations between quantitative variables

The main correlations between our quantitative features are:

- Age of driver is correlated with engine capacity
- Engine capacity is correlated with age of vehicle
- Number of vehicles is correlated with the number of casualties



4.1.5. Cramer's V for categorical features

After calculating the cramer's V for all the categorical features two by two, we took the variables with the biggest values of cramer's V with the target variable the created a heat map between these variables.

The categorical features kept for modelling after selection by cramer's V are:

vehicle_type, vehicle_leaving_carriageway, local_authority_district, urban_or_rural_area, and junction_control.

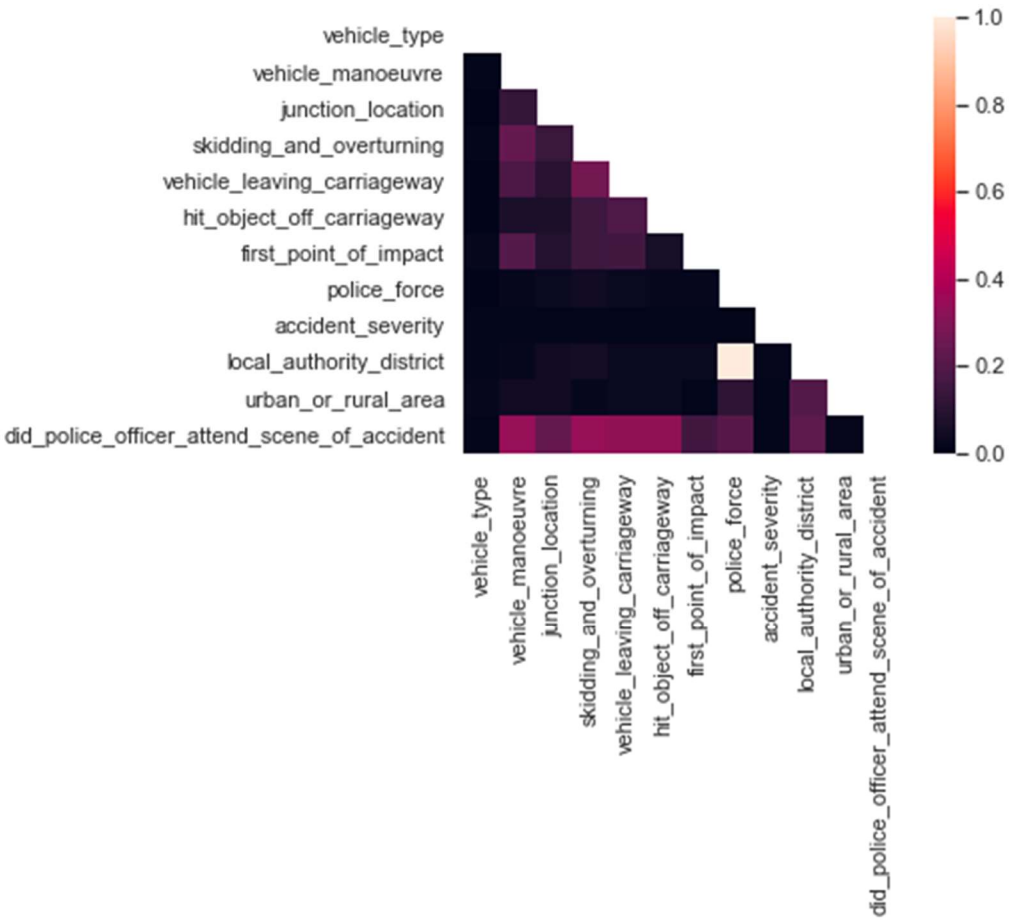


Figure 18 Cramer's V heat map

4.2. Model outputs

In order to train the model, we use 2019 data since it is the closest one to today's environment.

In order to get the most appropriate idea on the performance of the models, a 10 fold cross validation was conducted with the f1 score as the performance metric, the mean and standard deviation were printed as well as a boxplot showing the performance of each model.

The best models we got from the results are Random forest and Ada Boost. In order to choose between the two, a further investigation seems necessary.

```
kNN: 0.797292 (0.002900)
Naive Bayes: 0.783940 (0.003811)
DT: 0.701794 (0.003418)
RF: 0.804933 (0.002909)
AdaB: 0.803909 (0.002433)
```

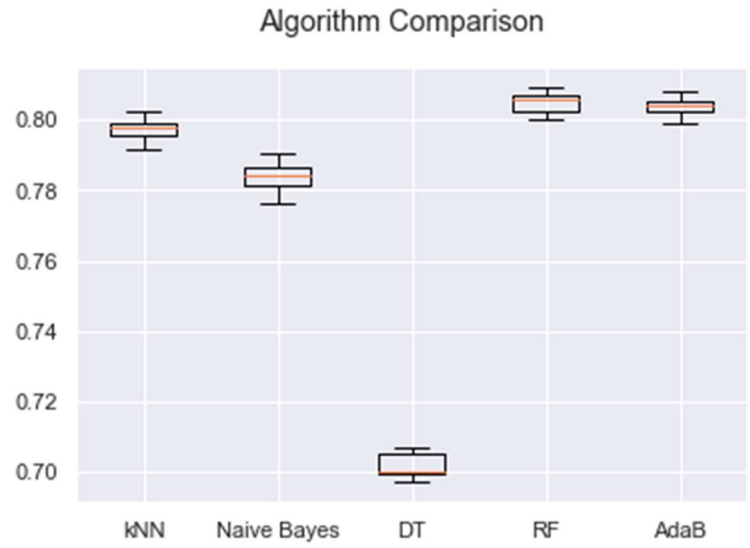


Figure 19 Models performance boxplot

The Random forest classifier performs better than Ada Boost classifier on predicting the fatal and the serious traffic accidents. However, it performs poorly on predicting the slight traffic accidents. (figure 21 and figure 22)

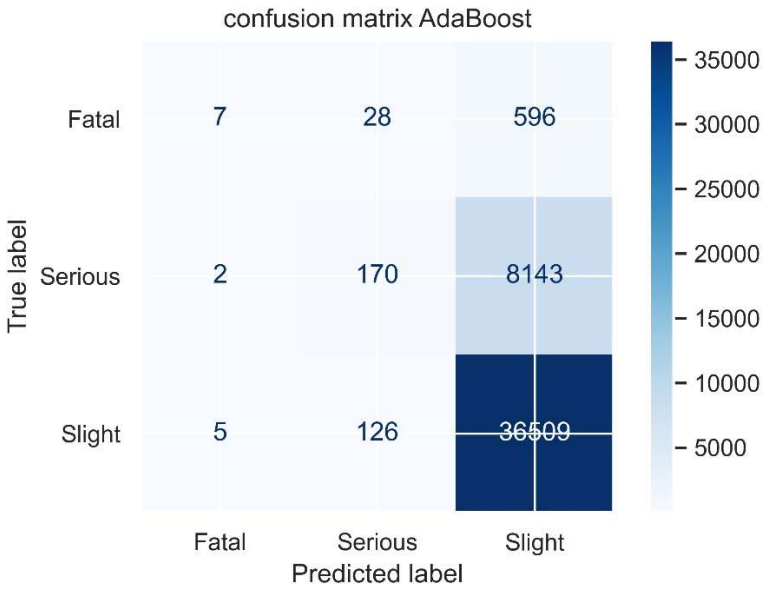


Figure 20 confusion matrix Ada Boost

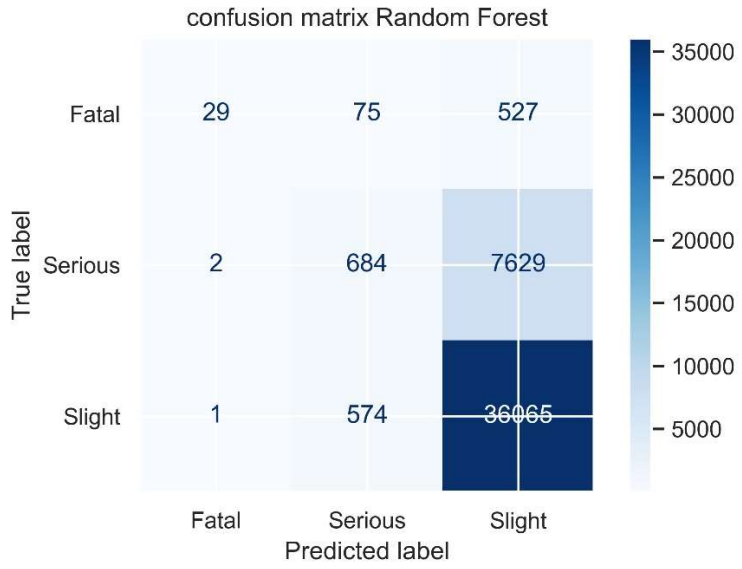


Figure 21 confusion matrix Random forest

The F1 score, the precision and the recall applied on the test set shows that the random forest classifier is performing better on the test set than the Adaboost classifier, since the F1 score for the random forest classifier is 0.7440, in contrast, the F1 score for the Ada Boost classifier is 0.7241.

	precision	recall	f1-score	support
class 0	0.50	0.01	0.02	631
class 1	0.52	0.02	0.04	8315
class 2	0.81	1.00	0.89	36640
accuracy			0.80	45586
macro avg	0.61	0.34	0.32	45586
weighted avg	0.75	0.80	0.72	45586

Figure 22 classification report Ada Boost

	precision	recall	f1-score	support
class 0	0.91	0.05	0.09	631
class 1	0.51	0.08	0.14	8315
class 2	0.82	0.98	0.89	36640
accuracy			0.81	45586
macro avg	0.74	0.37	0.37	45586
weighted avg	0.76	0.81	0.74	45586

Figure 23 Classification report Random Forest

For the feature importance, Ada Boost used vehicle type as the most important feature for classifying traffic accident severity then local authority district, which specifies the location, in the second place. However, Random Forest saw that local authority district comes in the first place then age of the driver in the second.

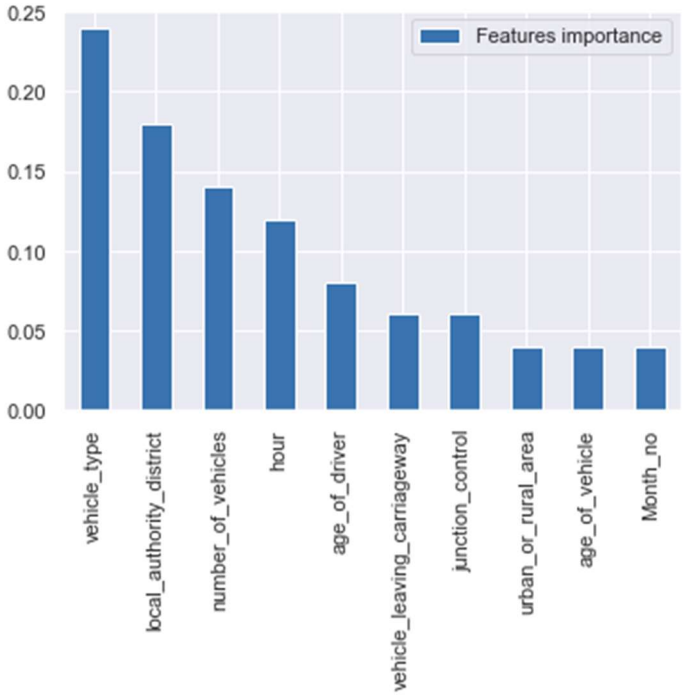


Figure 24 Feature importance Ada Boost

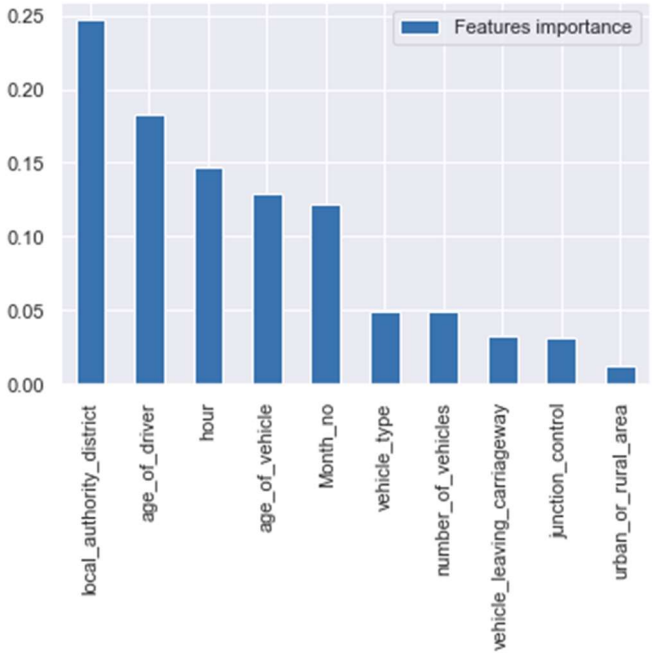


Figure 25 Feature importance Random Forest

To conclude, Random Forest is the classification model that performs the best overall. However, will it be better to create a pipeline that takes the output of the three best models, Random Forest, Ada Boost and K nearest neighbour, then apply a voting between the three with a preference for the Random Forest since it is the best model based on its performance.

5. Evaluation and discussion

5.1. EDA

In the exploratory data analysis conducted in 4.1., we have seen that the majority of the instances on our dataset are slight traffic accidents. The further analysis showed that, while the trend of the number of traffic accidents fluctuated decreasingly over time then dropped in 2020 because of COVID, the fatal traffic accidents did not decrease as much.

When investigating the number of accidents occurring by day of the week, we find that, as shown on the table 5, the number of accidents increases during the working days of the week, hitting its maximum point on Friday, then decreases dramatically on the weekend and hits the lowest point on Sunday. Weekday accident rates are significantly higher than weekend accident rates. Due to the distinct differences between the geographical effects of factors like subways, schools, and hospitals on weekdays and weekends.

However, the number of fatal accidents does not follow the same trend as the maximum points meets in the weekend, and fluctuates over the week. Where the other two categories follow the same trend. This is probably due to the increase of relentless driving during the weekend, which is amplified by the consumption of drugs, alcohol, and other substances

When we investigate the distribution of the number of accidents by hour on the weekdays, we find that during the working days, the distribution is slightly the same with the peak points at 8 am and 5 pm. However, on the weekends the distribution is different, where the peak points are between 12am and 6 pm.

When investigating the hour distribution on the three categories of the accident severity (slight, serious and fatal), we find that their distribution is nearly the same. With a difference on the fatal accidents, which does not have a peak point on 8am in contrary of serious and slight accidents, but they all have a peak point between 4pm and 5pm.

Most of the fatal accidents happen in rural areas 58.7 %, contrary to the serious and slight accidents, which mostly happen in urban areas, 55.8 % and 64.4 %. This is probably due to the fact that in and around areas the speed limit is high and the control measures are rare which leads to irresponsible driving and speed excess.

Most of the fatal traffic accidents happen on the roads with the speed limit of 30 and 60 miles per hour (33.76 % and 33.55 % respectively), when serious and slight ones mostly happen on the roads with a speed limit of 30 miles per hour.

The majority of traffic accidents happen on a fine without high winds weather, since most of the drivers take extra precaution when it is foggy or raining and drives slower with more attention taking into account the risk.

The drivers aged between 26 and 35 are most likely to be involved in a road accident, and the band between 36 and 45 comes in the second place. This is probably because of lack of experience and or the consumption of drugs and alcohol while driving.

The majority of the casualties are drivers or riders. This trend is also seen when crossing the casualty severity by the casualty class. However, pedestrians shows a big percentage as serious casualties, which means that the authorities needs to take more measures in order to protect them.

To conclude, it will much better if the dataset contained the driver, pedestrian or the circumstances who or which caused the accident. This, will give us a better understanding of the situations and the measures necessary to take.

5.2. Machine learning models

The best models we got from the box plot in figure 19 are Random forest and Ada Boost. In order to choose between the two, a further investigation seemed necessary.

The Random forest classifier performs better than Ada Boost classifier on predicting the fatal and the serious traffic accidents. However, it performs poorly on predicting the slight traffic accidents

The F1 score, the precision and the recall applied on the test set shows that the random forest classifier is performing better on the test set than the Adaboost classifier, since the F1 score for the random forest classifier is 0.7440, in contrast, the F1 score for the Ada Boost classifier is 0.7241.

For the feature importance, Ada Boost used vehicle type as the most important feature for classifying traffic accident severity then local authority district, which specifies the location,

in the second place. However, Random Forest saw that local authority district comes in the first place then age of the driver in the second.

To conclude, Random Forest is the classification model that performs the best overall. However, will it be better to create a pipeline that takes the output of the three best models, Random Forest, Ada Boost and K nearest neighbour, then apply a voting between the three with a preference for the Random Forest since it is the best model based on its performance.

6. References

Dataset: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Road-traffic injuries: confronting disparities to address a global-health problem, Shanthi Ameratunga, Martha Hijar, Robyn Norton 2006

Global road traffic injury statistics: Challenges, mechanisms and solutions. Fang-RongChang, He-Lai Huang, David C.Schwebel, Alan H.S.Chan, Guo-Qing Hu. 2020

Economics of Global Burden of Road Traffic Injuries and Their Relationship with Health System Variables. Koustuv Dalal, Zhiquin Lin, Mervyn Gifford, and Leif Svanström 2013. pages 1442–1450

Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. PLoS ONE 5, 1-9. Bettencourt LMA, Lobo J, Strumsky D, West GB. 2010

Uncovering the behaviour of road accidents in urban areas C. Cabrera-Arnau, R. Prieto Curiel and S. R. Bishop 2020

Tibebe Beshah Tesema, Ajith Abraham and Dejene Ejigu, "Learning the Classification of Traffic Accident Types", 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems, pp. 463-468, Sept. 2012.

Tibebe Beshah Tesema, Dejene Eijigu and Ajith Abraham, "Knowledge Discovery from Road Traffic Accident Data In Ethiopia", 2011 World Congress on Information and Communication Technologies, pp. 1241-1246, 2011.

Girija Narasimhan, Ben George Ben Ephrem et al., "Predictive Analytics of Road Accidents in Oman using Machine Learning Approach", 2017 International Conference on Intelligent Computing Instrumentation and Control technologies (ICICICT), pp. 1058-1065, July 2017.

L Li, S Shrestha and G Hu, "Analysis of Road Traffic Fatal Accidents using Data Mining Techniques", 2017 IEEE 15th International Conference on Software Engineering Research Management and Applications (SERA), pp. 363-370, 2017.

R Nidhi and V Kanchana, "Analysis of Road Accidents Using Data Mining Techniques", International Journal of Engineering & Technology, vol. 7, no. 3.10, pp. 40-44, 2018.

Classification of Road Traffic Accident Data Using Machine Learning Algorithms Bulbula Kumeda; Fengli Zhang; Fan Zhou; Sadiq Hussain; Ammar Almasri 2019

T. Beshah and S. Hill, "Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia", AAAI Spring Symposium, 2010.

G. Chen, Z. Zhang, R. Qian, R. A. Tarefder and Z. Tian, "Investigating Driver Injury Severity Patterns in Rollover Crashes Using Support Vector Machine Models", Accident Analysis and Prevention, vol. 90, pp. 128-139, 2016.

C Dong, C Shao, J Li and Z Xiong, "An improved deep learning model for traffic crash prediction", Journal of Advanced Transportation, pp. 1-13, 2018.

“Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity”, Rabia Emhamed AlMamlook; Keneth Morgan Kwayu; Maha Reda Alkasisbeh; 2019

Mandated societal lockdown and road traffic accidents Adnan I.Qureshi, Wei Huang, Suleman Khan and Iryna Lobanova 2020

Rokach, Lior, Maimon, Data mining with decision trees: theory and applications., 2nd Ed, World Scientific Pub Co Inc., 2015.

Quinlan, Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers 1986.

Hastie, Tibshirani, Friedman, The elements of statistical learning: data mining, inference, and prediction. New York: Springer Verlag, 2009.

Breiman, Friedman, Olshen, Stone, Classification and regression trees. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books 1984.

Roman Timofeev, Classification and Regression Trees (CART) Theory and Applications, Master Thesis, Université Humbold, Berlin, 2004.

Cornuejols, A., Miclet, L., Apprentissage Artificiel, Concepts et Algorithmes, Eyrolles, 2010.

7. Appendices

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 640331 entries, 0 to 42357
Data columns (total 33 columns):
 #   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
 0   accident_index                              640331 non-null  object
 1   accident_year                               640331 non-null  int64
 2   accident_reference                          640331 non-null  object
 3   police_force                                640331 non-null  int64
 4   accident_severity                           640331 non-null  int64
 5   number_of_vehicles                          640331 non-null  int64
 6   number_of_casualties                        640331 non-null  int64
 7   date                                        640331 non-null  object
 8   day_of_week                                 640331 non-null  int64
 9   time                                        640331 non-null  object
 10  local_authority_district                    640331 non-null  int64
 11  local_authority_ons_district                640331 non-null  object
 12  local_authority_highway                     640331 non-null  object
 13  first_road_class                            640331 non-null  int64
 14  first_road_number                           640331 non-null  int64
 15  road_type                                   640331 non-null  int64
 16  speed_limit                                 640294 non-null  float64
 17  junction_detail                             640331 non-null  int64
 18  junction_control                            640331 non-null  int64
 19  second_road_class                           640331 non-null  int64
 20  second_road_number                          640331 non-null  int64
 21  pedestrian_crossing_human_control           640331 non-null  int64
 22  pedestrian_crossing_physical_facilities     640331 non-null  int64
 23  light_conditions                            640331 non-null  int64
 24  weather_conditions                          640331 non-null  int64
 25  road_surface_conditions                     640331 non-null  int64
 26  special_conditions_at_site                  640331 non-null  int64
 27  carriageway_hazards                         640331 non-null  int64
 28  urban_or_rural_area                         640331 non-null  int64
 29  did_police_officer_attend_scene_of_accident 640331 non-null  int64
 30  trunk_road_flag                             640331 non-null  int64
 31  lsoa_of_accident_location                   640331 non-null  object
 32  Month                                       640331 non-null  object
dtypes: float64(1), int64(24), object(8)
memory usage: 166.1+ MB
```

Figure 26 The initial accident data informations

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 640331 entries, 0 to 42357
Data columns (total 35 columns):
 #   Column                                      Non-Null Count   Dtype
---  ------                                      --------------   -----
 0   accident_index                              640331 non-null  object
 1   accident_year                               640331 non-null  int64
 2   accident_reference                          640331 non-null  object
 3   police_force                                640331 non-null  category
 4   accident_severity                           640331 non-null  category
 5   number_of_vehicles                          640331 non-null  int64
 6   number_of_casualties                        640331 non-null  int64
 7   date                                        640331 non-null  datetime64[ns]
 8   day_of_week                                 640331 non-null  category
 9   time                                        640331 non-null  object
 10  local_authority_district                    640331 non-null  category
 11  local_authority_ons_district                640331 non-null  object
 12  local_authority_highway                     640331 non-null  object
 13  first_road_class                            640331 non-null  int64
 14  first_road_number                           640331 non-null  int64
 15  road_type                                   640331 non-null  category
 16  speed_limit                                 640331 non-null  category
 17  junction_detail                             640331 non-null  category
 18  junction_control                            640331 non-null  int64
 19  second_road_class                           640331 non-null  int64
 20  second_road_number                          640331 non-null  int64
 21  pedestrian_crossing_human_control           640331 non-null  category
 22  pedestrian_crossing_physical_facilities     640331 non-null  category
 23  light_conditions                            640331 non-null  category
 24  weather_conditions                          640331 non-null  category
 25  road_surface_conditions                     640331 non-null  category
 26  special_conditions_at_site                  640331 non-null  category
 27  carriageway_hazards                         640331 non-null  category
 28  urban_or_rural_area                         640331 non-null  category
 29  did_police_officer_attend_scene_of_accident 640331 non-null  category
 30  trunk_road_flag                             640331 non-null  category
 31  lsoa_of_accident_location                   640331 non-null  object
 32  Month                                       640331 non-null  object
 33  hour                                        640331 non-null  int64
 34  Month_no                                    640331 non-null  object
dtypes: category(17), datetime64[ns](1), int64(9), object(8)
memory usage: 103.8+ MB
```

Figure 27 The information of accident data after recoding

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1180002 entries, 0 to 78410
Data columns (total 27 columns):
 #   Column                              Non-Null Count    Dtype
---  ------                              --------------    -----
 0   accident_index                      1180002 non-null  object
 1   accident_year                       1180002 non-null  int64
 2   accident_reference                  1180002 non-null  object
 3   vehicle_reference                   1180002 non-null  int64
 4   vehicle_type                        1180002 non-null  category
 5   towing_and_articulation             1180002 non-null  category
 6   vehicle_manoeuvre                   1180002 non-null  category
 7   vehicle_direction_from              1180002 non-null  category
 8   vehicle_direction_to                1180002 non-null  category
 9   vehicle_location_restricted_lane    1180002 non-null  category
 10  junction_location                   1180002 non-null  category
 11  skidding_and_overturning            1180002 non-null  category
 12  hit_object_in_carriageway           1180002 non-null  category
 13  vehicle_leaving_carriageway         1180002 non-null  category
 14  hit_object_off_carriageway          1180002 non-null  category
 15  first_point_of_impact               1180002 non-null  category
 16  vehicle_left_hand_drive             1180002 non-null  category
 17  journey_purpose_of_driver           1180002 non-null  category
 18  sex_of_driver                       1180002 non-null  category
 19  age_of_driver                       1180002 non-null  int64
 20  age_band_of_driver                  1180002 non-null  category
 21  engine_capacity_cc                  1180002 non-null  int64
 22  propulsion_code                     1180002 non-null  category
 23  age_of_vehicle                      1180002 non-null  int64
 24  generic_make_model                  1180002 non-null  object
 25  driver_imd_decile                   1180002 non-null  category
 26  driver_home_area_type               1180002 non-null  category
dtypes: category(19), int64(5), object(3)
memory usage: 102.4+ MB
```

Figure 28 The initial vehicle data informations

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1180002 entries, 0 to 78410
Data columns (total 27 columns):
 #   Column                          Non-Null Count    Dtype
---  ------                          --------------    -----
 0   accident_index                  1180002 non-null  object
 1   accident_year                   1180002 non-null  int64
 2   accident_reference              1180002 non-null  object
 3   vehicle_reference               1180002 non-null  int64
 4   vehicle_type                    1180002 non-null  category
 5   towing_and_articulation         1180002 non-null  category
 6   vehicle_manoeuvre               1180002 non-null  category
 7   vehicle_direction_from          1180002 non-null  category
 8   vehicle_direction_to            1180002 non-null  category
 9   vehicle_location_restricted_lane 1180002 non-null  category
 10  junction_location               1180002 non-null  category
 11  skidding_and_overturning        1180002 non-null  category
 12  hit_object_in_carriageway       1180002 non-null  category
 13  vehicle_leaving_carriageway     1180002 non-null  category
 14  hit_object_off_carriageway      1180002 non-null  category
 15  first_point_of_impact           1180002 non-null  category
 16  vehicle_left_hand_drive         1180002 non-null  category
 17  journey_purpose_of_driver       1180002 non-null  category
 18  sex_of_driver                   1180002 non-null  category
 19  age_of_driver                   1180002 non-null  int64
 20  age_band_of_driver              1180002 non-null  category
 21  engine_capacity_cc              1180002 non-null  int64
 22  propulsion_code                 1180002 non-null  category
 23  age_of_vehicle                  1180002 non-null  int64
 24  generic_make_model              1180002 non-null  object
 25  driver_imd_decile               1180002 non-null  category
 26  driver_home_area_type           1180002 non-null  category
dtypes: category(19), int64(5), object(3)
memory usage: 102.4+ MB
```

Figure 29 Vehicle data after recoding informations

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 834890 entries, 0 to 53173
Data columns (total 19 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   accident_index                        834890 non-null  object
 1   accident_year                         834890 non-null  int64
 2   accident_reference                    834890 non-null  object
 3   vehicle_reference                     834890 non-null  int64
 4   casualty_reference                    834890 non-null  int64
 5   casualty_class                        834890 non-null  int64
 6   sex_of_casualty                       834890 non-null  int64
 7   age_of_casualty                       834890 non-null  int64
 8   age_band_of_casualty                  834890 non-null  int64
 9   casualty_severity                     834890 non-null  int64
 10  pedestrian_location                   834890 non-null  int64
 11  pedestrian_movement                   834890 non-null  int64
 12  car_passenger                         834890 non-null  int64
 13  bus_or_coach_passenger                834890 non-null  int64
 14  pedestrian_road_maintenance_worker    834890 non-null  int64
 15  casualty_type                         834890 non-null  int64
 16  casualty_home_area_type               834890 non-null  int64
 17  casualty_imd_decile                   834890 non-null  int64
 18  status                                53174 non-null   object
dtypes: int64(16), object(3)
memory usage: 127.4+ MB
```

Figure 30 initial casualty data informations

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 834890 entries, 0 to 53173
Data columns (total 18 columns):
 #   Column                                Non-Null Count   Dtype
---  ------                                --------------   -----
 0   accident_index                        834890 non-null  object
 1   accident_year                         834890 non-null  int64
 2   accident_reference                    834890 non-null  object
 3   vehicle_reference                     834890 non-null  int64
 4   casualty_reference                    834890 non-null  int64
 5   casualty_class                        834890 non-null  category
 6   sex_of_casualty                       834890 non-null  category
 7   age_of_casualty                       834890 non-null  int64
 8   age_band_of_casualty                  834890 non-null  category
 9   casualty_severity                     834890 non-null  category
 10  pedestrian_location                   834890 non-null  category
 11  pedestrian_movement                   834890 non-null  category
 12  car_passenger                         834890 non-null  category
 13  bus_or_coach_passenger                834890 non-null  category
 14  pedestrian_road_maintenance_worker    834890 non-null  category
 15  casualty_type                         834890 non-null  category
 16  casualty_home_area_type               834890 non-null  category
 17  casualty_imd_decile                   834890 non-null  category
dtypes: category(12), int64(4), object(2)
memory usage: 54.1+ MB
```

Figure 31 Casualty data after recoding informations