Predicting the Severity of Road Accidents

Coursera and IBM Data Science Professional Certificate Capstone Project

By Liis Monson | Sept 2020

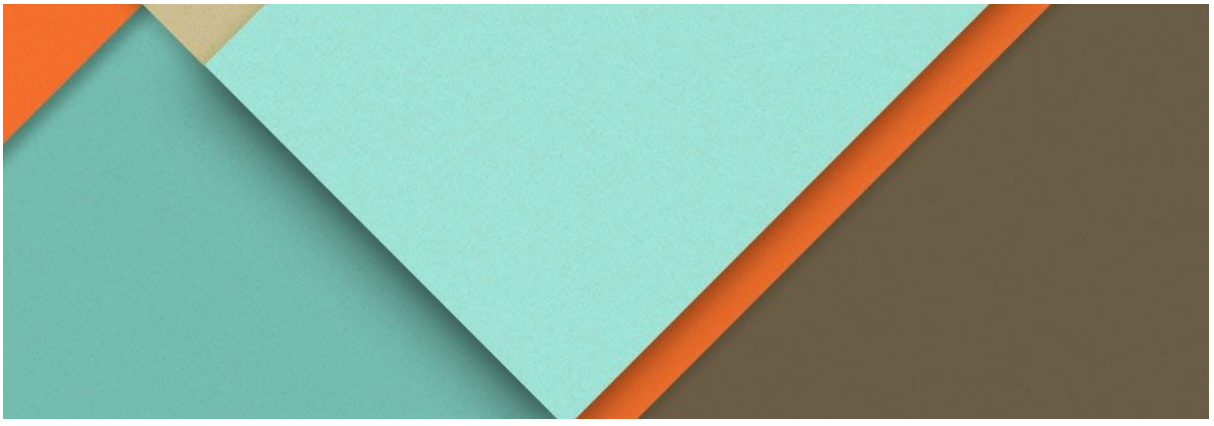


Table of Contents

1 Introduction and Business Understanding

2 Data Understanding

3 Methodology

- 3.1 Data Preparation & Exploratory Analysis
  - 3.1.1 Casualty Severity Assessment
    - 3.1.1.1 Casualty Severity
    - 3.1.1.2 Number of Vehicles
    - 3.1.1.3 Number of Casualties per Accident
  - 3.1.2 Severity Score
    - 3.1.2.1 Severity Score Calculation
    - 3.1.2.2. SS Categorisation
  - 3.1.3 Environmental Features
    - 3.1.3.1 Time of Day
    - 3.1.3.2 Seasonality
    - 3.1.3.3 1st Road Class
    - 3.1.3.4 Road Surface Conditions
    - 3.1.3.5 Lighting Conditions
    - 3.1.3.6 Weather Conditions
  - 3.1.4 Casualty's Personal Features
    - 3.1.4.1 Type of Vehicle
    - 3.1.4.2 Age and Gender of Casualties
  - 3.1.5 Final Data Set
- 3.2 Data Modelling
  - 3.2.1 K-Nearest Neighbor KNN
  - 3.2.2 Decision Tree

4 Model Evaluation & Results

- 4.1 Results

5 Discussion & Conclusion

1 Introduction and Business Understanding

This project is done according to CRISP-DM process aka “Cross-Industry Standard Process for Data Mining”.

Technical details of this work can be found in my GitHub Repository.

Road accidents cause every year a significant amount of injuries, death and damage to the property. Providing an accurate prediction of the accident severity can be extremely valuable to all the emergency responders like ambulance teams, ER departments, traffic police, and rescue services. It could be a tool to estimate the resources needed, potential impact of an accident, as well as efficiently execute the emergency management procedures. The immediate aftermath and the emergency services’ reaction to an accident are crucial elements to deal with the accidents the most efficiently, and therefore reduce the amount of injuries, complications, impact, cost, and save lives.

In order to be able to assess the severity of an accident, we should find answers to the following questions:

- How do we define the severity of an accident?
- Is there a clear correlation between road accident severity rate and environmental factors like time of a day, light, weather and road conditions?
- Is there a clear correlation between accident severity rate and driver’s characteristics like age, gender, and type of vehicle?
- Is there a clear correlation between driver’s characteristics, environmental conditions and severity of occured accidents?
- Can we predict the accident severity rate with relatively high probability using the insights above?

2 Data Understanding

In order to answer the questions above, the data set chosen has to include environmental features, drivers' features as well as features needed for the severity assessment of the accidents.

The used data set has been chosen from UK Government Digital Service's Data site data.gov.uk and represents accident data from Leeds for 2018.

This data was chosen for the following reasons:

- It is easy accessible public data from a trusted source
- Data is available for several years, and therefore the model can be trained and then tested on out-of-sample data set
- Data includes the necessarily features for answering the questions asked in paragraph 1, including environmental observations, date and time, drivers' characteristics, vehicle type and severity rate
- Data sets are big enough for modelling, but small enough to be able to work comfortably from home PC
- Data is already well formatted which will cut down the processing time. It already contains mainly numerical values that can be used for statistical analysis, and the values are clearly organised into feature columns.

Characteristics of the original data set:

- 1995 rows. Each row represents a person involved in an accident.
- 21 columns. Column names, examples, data types and descriptions:

Column Name	Example	Dtype	Description
Accident Fields_Reference Number	51B0230	Object	Accident's ref number
Grid Ref: Easting	433936	Int64	Eastward coordinate of the location
Grid Ref: Northing	428874	Int	Northward coordinate of the location
Number of Vehicles	1	Int	Number of vehicles involved in accident
Accident Date	11/01/2018	Object	Accident date in DD/MM/YYYY format
Time (24hr)	700	Int64	Time in 24h format
1st Road Class	6	Int64	Road class
1st Road Class & No	U	Object	Road number
Road Surface	1	Int64	Road surface description

Lighting Conditions	4	Int64	Lighting condititon descriptions
Weather Conditions	1	Int64	Weather condition descriptions
Local Authority	E08000035	Object	Ref of local authority dealing with accident
Vehicle Fields_Reference Number	51B0230	Object	Vehicle reference number for local authorities
Vehicle Number	1	Int64	The licence plate number that has been removed form data set and replaced with 1
Type of Vehicle	9	Int64	The type of vehicle that caused the accident
Casualty Fields_Reference Number	51B0230	Object	Reference number of the victim
Casualty Veh No	1	Int64	Number of the casualty's vehicle involved in accident (e.g. 1st, 2nd 3rd vehicle out of 3)
Casualty Class	3	Int64	Classifies if the casualty was driver, passenger or pedestrian
Casualty Severity	3	Int64	Severity of injuries: 3 is slight, 1 is fatal
Sex of Casualty	2	Int64	Casualty's gender
Age of Casualty	87	Int64	Casualty's age in numbers

- Descriptions of numeric values within the data set:

Feature Name	**Reference Number**	**Description**
1st Road Class	1	Motorway
1st Road Class	2	A(M)
1st Road Class	3	A
1st Road Class	4	B
1st Road Class	5	C
1st Road Class	6	Unclassified
Road Surface	1	Dry
Road Surface	2	Wet / Damp
Road Surface	3	Snow
Road Surface	4	Frost / Ice
Road Surface	5	Flood (surface water over 3cm deep)
Lighting Conditions	1	Daylight: street lights present

Lighting Conditions	2	Daylight: no street lighting
Lighting Conditions	3	Daylight: street lighting unknown
Lighting Conditions	4	Darkness: street lights present and lit
Lighting Conditions	5	Darkness: street lights present but unlit
Lighting Conditions	6	Darkness: no street lighting
Lighting Conditions	7	Darkness: street lighting unknown
Weather Conditions	1	Fine without high winds
Weather Conditions	2	Raining without high winds
Weather Conditions	3	Snowing without high winds
Weather Conditions	4	Fine with high winds
Weather Conditions	5	Raining with high winds
Weather Conditions	6	Snowing with high winds
Weather Conditions	7	Fog or mist – if hazard
Weather Conditions	8	Other
Weather Conditions	9	Unknown
Casualty Class	1	Driver or rider
Casualty Class	2	Vehicle or pillion passenger
Casualty Class	3	Pedestrian
Casualty Severity	1	Fatal
Casualty Severity	2	Serious
Casualty Severity	3	Slight
Sex of Casualty	1	Male
Sex of Casualty	2	Female
Type of Vehicle	1	Pedal cycle
Type of Vehicle	2	M/cycle 50cc and under
Type of Vehicle	3	Motorcycle over 50cc and up to 125cc
Type of Vehicle	4	Motorcycle over 125cc and up to 500cc
Type of Vehicle	5	Motorcycle over 500cc
Type of Vehicle	6	[Not used]
Type of Vehicle	7	[Not used]
Type of Vehicle	8	Taxi/Private hire car
Type of Vehicle	9	Car
Type of Vehicle	10	Minibus (8 – 16 passenger seats)
Type of Vehicle	11	Bus or coach (17 or more passenger seats)
Type of Vehicle	12	[Not used]
Type of Vehicle	13	[Not used]
Type of Vehicle	14	Other motor vehicle
Type of Vehicle	15	Other non-motor vehicle
Type of Vehicle	16	Ridden horse

Type of Vehicle	17	Agricultural vehicle (includes diggers etc.)
Type of Vehicle	18	Tram / Light rail
Type of Vehicle	19	Goods vehicle 3.5 tonnes mgw and under
Type of Vehicle	20	Goods vehicle over 3.5 tonnes and under 7.5 tonnes mgw
Type of Vehicle	21	Goods vehicle 7.5 tonnes mgw and over
Type of Vehicle	22	Mobility Scooter
Type of Vehicle	90	Other Vehicle
Type of Vehicle	97	Motorcycle - Unknown CC

Insights to data from df.describe() function:

	Accident Fields_Reference Number	Number of Vehicles	Accident Date	Time (24hr)	Age of Casualty
count	1995	1995.000000	1995	1995.000000	1995.000000
unique	1548	NaN	354	NaN	NaN
top	58K1318	NaN	30/06/2018	NaN	NaN
freq	9	NaN	19	NaN	NaN
mean	NaN	1.919298	NaN	1409.368421	36.993985
std	NaN	0.723584	NaN	505.175755	18.856635
min	NaN	1.000000	NaN	0.000000	1.000000
25%	NaN	1.000000	NaN	1047.500000	23.000000
50%	NaN	2.000000	NaN	1503.000000	35.000000
75%	NaN	2.000000	NaN	1754.000000	50.000000
max	NaN	7.000000	NaN	2357.000000	95.000000

- There are 1548 unique numbers in Accident Fields_Reference Number, meaning the 1995 casualties were result of 1548 accidents making the average number of casualties 1.29 per accident
- The highest number of casualties per accident was 9
- Most of the accidents have 2 vehicle involved, the max amount of vehicles involved was 7
- There are 354 unique values for dates, meaning there were 11 days were no accidents happened
- The highest number of casualties was 19, this happened on 30th June 2018
- Most of the injuries happen during the day, highest correlation seems to be with afternoon rush hour as ~25% casualties happened between 3.05PM and 5.54PM
- More accidents happen with younger people because 50% of casualties were =<35 years old

The features from the data set that could be potentially used to achieve the objective:

- Environmental data:

    - Accident Date - potentially seasonality might be a factor
    - Accident Time - time of the day might be a factor
    - 1st Road Class
    - Road Surface
    - Lighting Conditions
    - Weather Conditions
- Casualty data:
    - Type of Vehicle
    - Sex of Casualty
    - Age of Casualty
- Features for assessing the severity of accident:
    - Accident Field Reference number - more rows i.e. persons are marked with the same number, the more casualties the accident had
    - Number of Vehicles - more vehicles involved in accident, the more severe the accident
    - Casualty Severity - the extent of injuries

At this stage all the columns that are not listed were dropped from the data set in order to simplify the data. These were the following: 'Grid Ref: Easting', 'Grid Ref: Northing', '1st Road Class & No', 'Local Authority', 'Vehicle Fields_Reference Number', 'Vehicle Number', 'Casualty Fields_Reference Number', 'Casualty Veh No', 'Casualty Class'.

Renamed Accident Fields_Reference Number to Accident Ref No for simplicity

That's how the data frame looks now:

	Accident Ref No	Number of Vehicles	Accident Date	Time (24hr)	1st Road Class	Road Surface	Lighting Conditions	Weather Conditions	Type of Vehicle	Casualty Class	Casualty Severity	Sex of Casualty	Age of Casualty
0	51B0230	1	11/01/2018	700	6	1	4	1	9	3	3	2	87
1	51B0349	1	11/01/2018	855	6	2	1	1	11	2	3	2	60
2	51B0349	1	11/01/2018	855	6	2	1	1	11	2	3	2	51
3	51B0349	1	11/01/2018	855	6	2	1	1	11	2	3	2	50
4	51B0349	1	11/01/2018	855	6	2	1	1	11	2	3	2	51

3 Methodology

3.1 Data Preparation & Exploratory Analysis

Going through all the relevant features listed, plotting them and making sure they make sense in context of the objective.
As it is not yet known which ML model will be used, or is the most accurate, we need to make sure that the whole data is in a format that can be used for all the models. I.e. it has to have all the categorical data fields also available in numerical values.

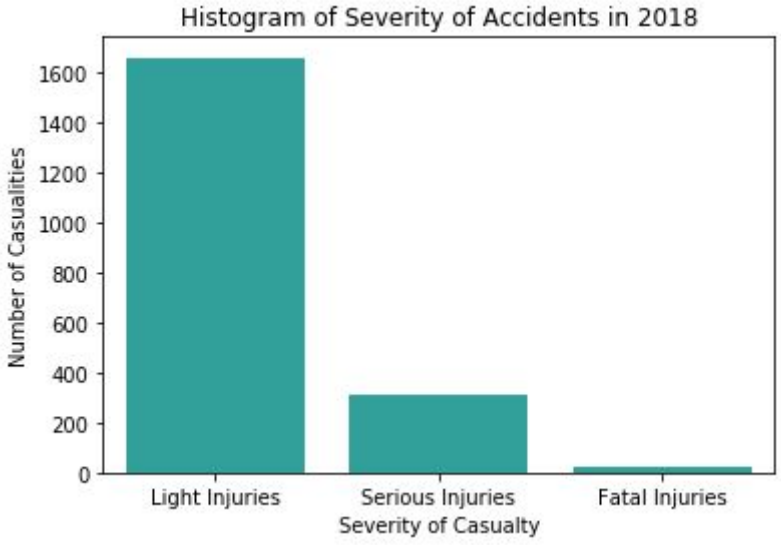
3.1.1 Casualty Severity Assessment

In order to be able to assess and categorise the accidents with machine learning model, we need to review the data and create the classes that then later could be predicted.

3.1.1.1 Casualty Severity

First we need to assess if the casualty severity data is usable, in a good format and diverse enough to draw conclusions.

Light Injuries	1658	83.1%
Serious Injuries	311	15.6%
Fatal Injuries	26	1.3%



Because the number of Fatal Injuries is very small, I merged the Fatal and Serious Injuries together and called the categorisation “Serious or Fatal Injury”.
I also changed the value in the Casualty Severity column to mark Light Injury with 1 and Serious or Fatal Injury with 2 because I find it more intuitive.

	Casualty Severity	Casualty Severity Description
1985	1	Light Injuries
1986	1	Light Injuries
1987	2	Serious or Fatal Injuries
1988	1	Light Injuries
1989	1	Light Injuries
1990	2	Serious or Fatal Injuries

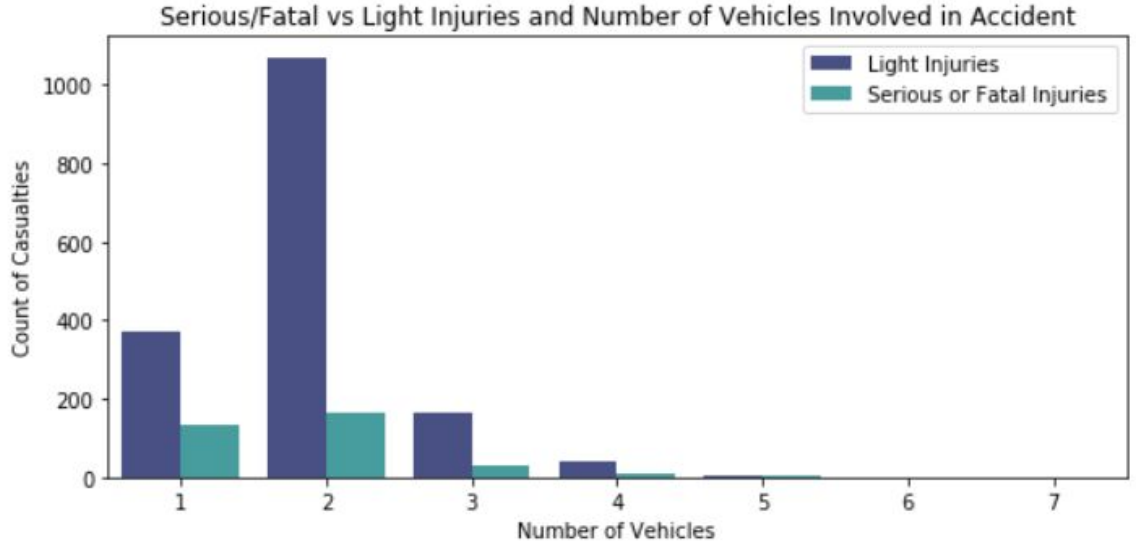
3.1.1.2 Number of Vehicles

How many vehicles were involved and how many casualties were caused by accident are definitely factors of assessing how severe the accident was.

First let's see how many vehicles as an average were involved in accidents:

Number of Vehicles	% of Casualties
1	0.251629
2	0.618546
3	0.097744
4	0.025564
5	0.005013
6	0.000501
7	0.001003

The division of serious and fatal vs light injuries per number of vehicles involved in accidents is shown below. It is clear that when there's only one vehicle involved, the chance of having a serious or fatal casualty is going up considerably.



Number of Vehicles	Casualty Severity Description	%
1	Light Injuries	0.739044

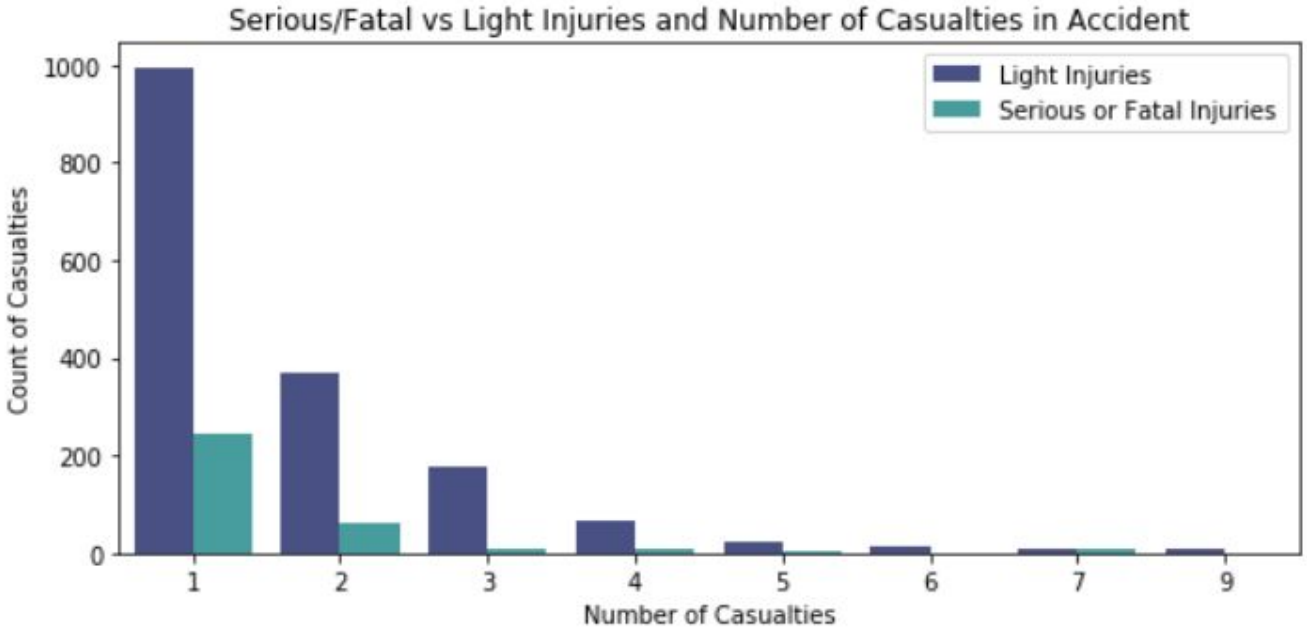
	Serious or Fatal Injuries	0.260956
2	Light Injuries	0.867909
	Serious or Fatal Injuries	0.132091
3	Light Injuries	0.851282
	Serious or Fatal Injuries	0.148718
4	Light Injuries	0.823529
	Serious or Fatal Injuries	0.176471
5	Light Injuries	0.600000
	Serious or Fatal Injuries	0.400000
6	Light Injuries	1.000000
7	Light Injuries	0.500000
	Serious or Fatal Injuries	0.500000

3.1.1.3 Number of Casualties per Accident

In order to calculate how many people were injured in any accident, I counted the identical reference numbers in column 'Accident Ref No' and entered the counts into the dataframe with a new column called 'Number of Casualties'.
The resulting reference system:

	Accident Ref No	Number of Casualties
0	51B0230	1
1	51B0349	5
2	51B0349	5
3	51B0349	5
4	51B0349	5
5	51B0349	5
6	51B0632	1
7	51B0975	1
8	51B1212	1
9	51B1241	1

Plotted the number to figure out how many casualties there normally are per accident, as well as how are the numbers divided in respect to the Casualty Severity.



It appears that most of the accidents involve just one casualty, in which case the chance that the casualty severity is higher is also greater. All in all it seems that more people are injured in an accident, the lower the chance of serious or fatal injuries.

Number of Casualties	Casualty Severity Description	%
1	Light Injuries	0.802419
	Serious or Fatal Injuries	0.197581
2	Light Injuries	0.854839
	Serious or Fatal Injuries	0.145161
3	Light Injuries	0.941799
	Serious or Fatal Injuries	0.058201
4	Light Injuries	0.902778
	Serious or Fatal Injuries	0.097222
5	Light Injuries	0.840000
	Serious or Fatal Injuries	0.160000
6	Light Injuries	1.000000
7	Light Injuries	0.500000
	Serious or Fatal Injuries	0.500000
9	Light Injuries	1.000000

3.1.2 Severity Score

Severity Score (SS) is a numeric value given to each row of data in the dataset to have a clear comparison and instant assessment of severity for each accident.

3.1.2.1 Severity Score Calculation

The features in dataset that are used to assess the severity of the accident:

- Casualty Severity / Casualty Severity Description - extent of personal injuries of each casualty
- Number of Vehicles - reflects damage to the property and potential disruption of traffic flow
- Number of Casualties - total number of people injured in the accident

All the features above were taken into account and the following algorithm was used to calculate the score for each accident:

$$SS = V + (CS1 + CS2 + ... + CSn)$$

Where:
SS = *Severity Score*
V = *Number of Vehicles*
CS = *Casualty Severity Value of each casualty*

Casualty Severity (CS) value for serious or fatal injury is 2, and therefore they have more weight than Number of Vehicles or light injuries with CS value 1.

The above score means that for example an accident which involves one vehicle and one light casualty would have SS = 1 + 1 = 2. An accident with two vehicles and two serious casualties would have SS = 2 + 2 + 2 = 6.

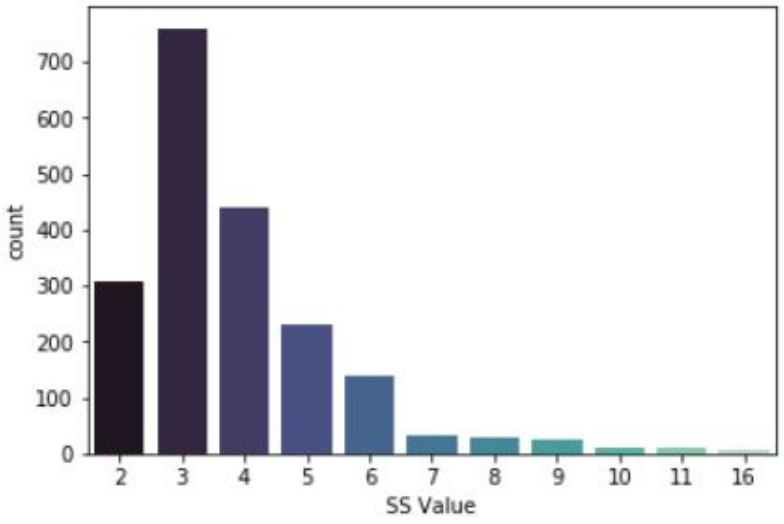
In order to add the SS value to the data frame, a new column was added for CS Value (short for Casualty Severity Value of the Accident) which represents the (CS1 + CS2+....+CSn) part of the formula above. It is a total Casualty Severity score per accident reflecting the severity scores of all casualties.

The next step was to sum the Number of Vehicles and the CS Values in order to get the overall SS value. The SS value was also added as a new column so each row has now the SS for an accident:

	Accident Ref No	Number of Vehicles	CS Value	SS Value
1985	5CR1051	2	1	3
1986	5CR1299	2	1	3
1987	5CU0470	2	2	4
1988	5CU0926	2	1	3
1989	5CU1803	2	1	3
1990	5CV0814	2	5	7
1991	5CV0814	2	5	7
1992	5CV0814	2	5	7
1993	5CV0814	2	5	7
1994	5CV1097	1	1	2

In the current data set the min SS value is 2 and the max SS value is 16:

SS Score	**Count**
2	309
3	760
4	441
5	231
6	139
7	34
8	28
9	25
10	12
11	9
16	7



The minimum value of SS with the current data set is 2 because each recorded casualty has at least light injuries (CS value = 1) and there was always at least 1 vehicle involved. However, the formula also would work if there was an accident involving a vehicle but no casualties, or involving a casualty without vehicle, which is good as therefore the formula could be applied outside of the current data borders.

The maximum value in the current data set is 16, that is caused by one severe accident with 2 vehicles and 7 severely injured casualties. In reality there is no max value of the SS, in case of catastrophe it could reach hundreds.

	Accident Ref No	Number of Vehicles	Time of day	Season	Road Surface Description	Lighting Description	Casualty Severity	Sex of Casualty	Age of Casualty	SS Value
1038	56U0266	2	Night	Summer	Dry	Dark, street lights	2	2	17	16
1039	56U0266	2	Night	Summer	Dry	Dark, street lights	2	2	17	16
1040	56U0266	2	Night	Summer	Dry	Dark, street lights	2	1	18	16
1041	56U0266	2	Night	Summer	Dry	Dark, street lights	2	1	19	16
1042	56U0266	2	Night	Summer	Dry	Dark, street lights	2	1	21	16
1043	56U0266	2	Night	Summer	Dry	Dark, street lights	2	1	19	16
1044	56U0266	2	Night	Summer	Dry	Dark, street lights	2	1	42	16

3.1.2.2. SS Categorisation

An assumption is made that an unlimited number of categories are too many to use effectively in daily life, and for a model to accurately categorise the data. Too little data and too many categories will inevitably lead to mis-categorisation. In the current data set the higher values are produced by factual outliers, and the size of the data set is not sufficiently large to establish if these outliers are representing an actual trend or are exceptions.

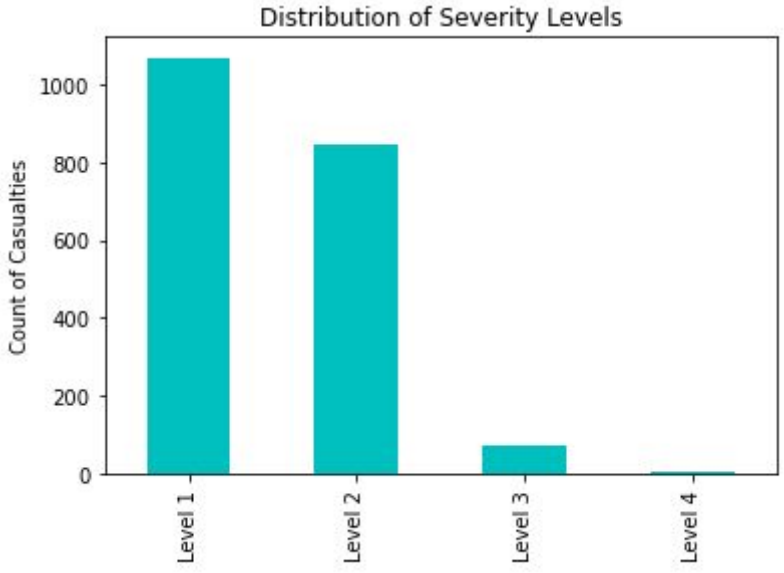
It would also be easier for emergency services to have less, but clearly labelled categories. The categories are increasing exponentially i.e. Severity Level 2 and Severity Level 4 have a difference of 2 levels and therefore the Level 4 accident is 2x2=4 times more serious than the Level 2 Accident.

The data was categorised according to the SS Values as following:

SS Value range	**Given Classification**	**Reasoning**
1	Severity Level 0	It's an incident without more serious consequences.
2-3	Severity Level 1	Serious incidents / smaller accidents. Max number of vehicles involved is 2, max number of serious injuries is 1.
4-7	Severity Level 2	Medium severity accidents with local consequences. Max number of vehicles involved is 6, max number of serious injuries is 3.
8-13	Severity Level 3	Serious accidents with either many vehicles and many casualties, or both. Max number of vehicles involved is 12, max number of serious injuries is 6.

14-25	Severity Level 4	Very serious accidents. Max number of vehicles involved is 24, max number of serious injuries is 12.
26-50	Severity Level 5	Major accidents. Max number of vehicles involved is 49, max number of serious injuries is 25.
41-...	Severity Level 6	A potential disaster situation.

I have added a new column called Severity Level to the data set and mapped all the rows:



3.1.3 Environmental Features

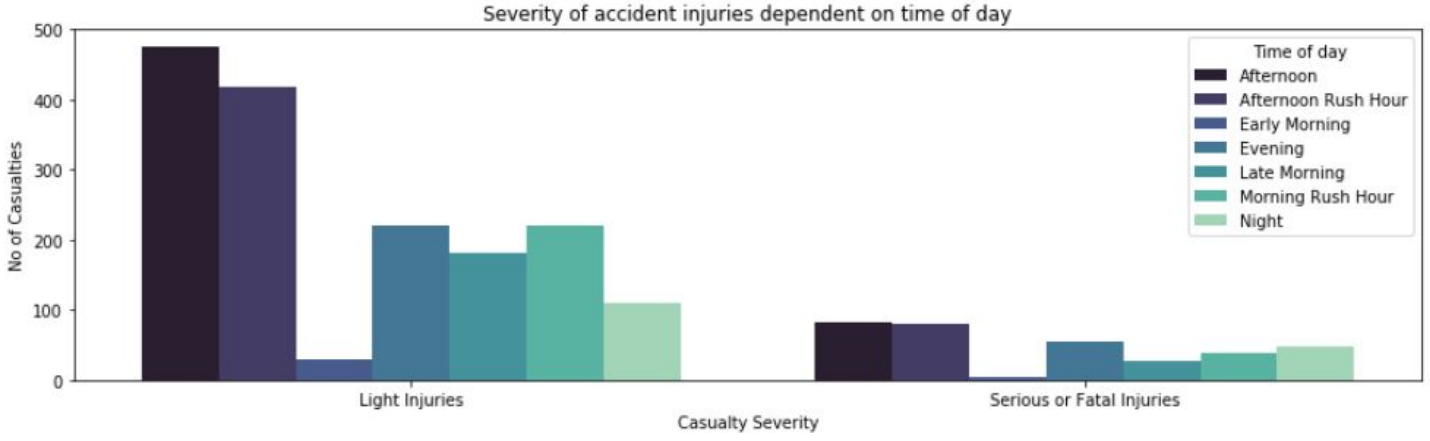
A feature by feature overview of all the features in the data set that are affected by the surrounding environment. The overview includes the original data, the processing and formatting that was done to the original data, and any exploratory analysis that was done to establish if it's a good feature to be used in the ML model.

3.1.3.1 Time of Day

The hypothesis would be that the time of day affects the possibility of getting into accidents as well as severity of accidents that happen.
In order to understand better if and how the time of day affects the severity and amount of accidents, I added a column 'Time of day' to the data set and divided the column 'Time (24h)' in to the following categories:

Time (24) column value range	Length of period	24h 12h description	Value for new 'Time of day' columns
> 2230 : <= 500	6.5h	From 22:30 - 05:00 From 10:30PM - 5AM	Night
>500 : <=700	2h	From 05:00 - 07:00 From 5AM - 7AM	Early Morning
>700 : <=930	2.5h	From 07:00 - 09:30 From 7AM - 9:30AM	Morning Rush Hour
>930 : <=1200	2.5h	From 09:30 - 12:00 From 9:30AM - 12PM	Late Morning
>1200 : <=1600	4h	From 12:00 - 16:00 From 12PM - 4PM	Afternoon
>1600 : <=1900	3h	From 16:00 - 19:00 From 4PM - 7PM	Afternoon Rush Hour
>1900 : <=2230	3.5h	From 19:00 - 22:30 From 7PM - 10:30PM	Evening

Because the number of Fatal Injuries is very small, I merged the Fatal and Serious Injuries together and called the categorisation "Serious or Fatal Injury".
I also changed the value in the Casualty Severity column to mark Light Injury with 1 and Serious or Fatal Injury with 2 because I find it more intuitive.
Then I took these values and plotted them on bar chart together with the casualty severity parameter:



We can clearly see that most of the casualties happen in the afternoon and afternoon rush hour, that most probably can be explained by the increased volume of traffic as well as general fatigue after working day. However, the number of accidents is high, but the overall severity of the accidents is lower than expected (see below in *italic*).

The overall possibility of having an accident during the night goes down, however the possibility of having serious accident goes up considerably (**bold** below):

Overall division of casualties based on time of day:		**%**
	Afternoon	0.280201
	Afternoon Rush Hour	0.250125
	Evening	0.138346
	Morning Rush Hour	0.130326
	Late Morning	0.104261
	Night	0.079198
	Early Morning	0.017544

Division of casualties based on time of day and accident severity:

Casualty Severity	**Time of day**	**%**
Light Injuries	*Afternoon*	*0.287093*
	Afternoon Rush Hour	*0.252714*
	Evening	0.133293
	Morning Rush Hour	0.133293
	Late Morning	0.109168
	Night	**0.066345**
	Early Morning	0.018094
Serious or Fatal Injuries	*Afternoon*	*0.246291*
	Afternoon Rush Hour	*0.237389*
	Evening	0.163205
	Night	**0.142433**
	Morning Rush Hour	0.115727
	Late Morning	0.080119
	Early Morning	0.014837

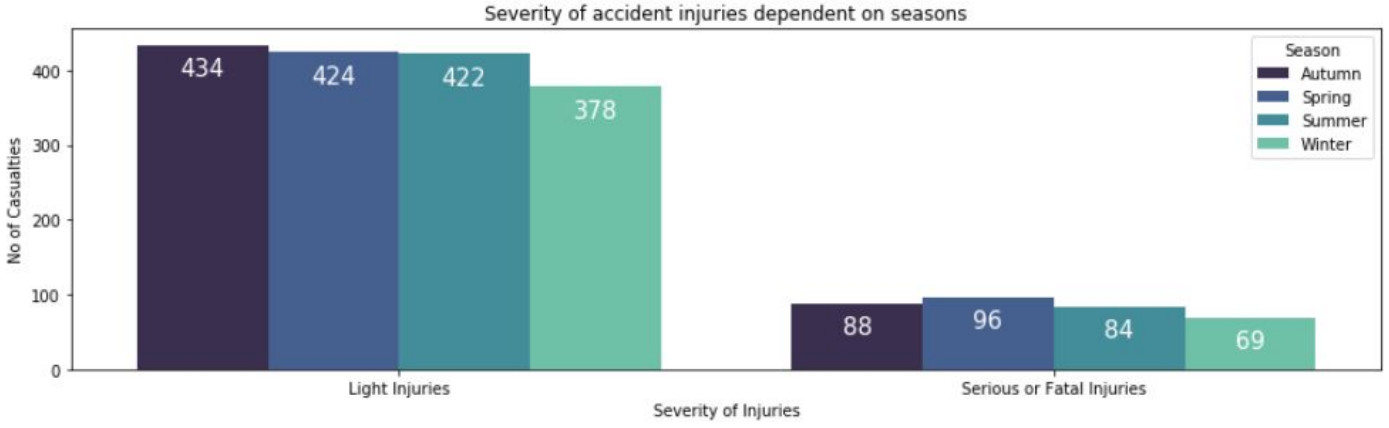
Time of day will have to be counted for the accident severity and possibility rate predictions as it seems that during the afternoon and afternoon rush hour one is more likely to have an accident, and during the night one is more likely to have a serious accident.

3.1.3.2 Seasonality

There is a good chance that not only time of day but also seasonality might affect the accident rate as well as casualty severity.
In order to have a better grasp of data, I added a new column called 'Season' to the dataframe based on the Accident Date column and divided it as following:

Date range	Season
1st December - 28th February	Winter
1st March - 31st May	Spring
1st June - 31st August	Summer
1st September - 30th November	Autumn

Ran a bar chart on the data, from where we can see that there seems to be proportionally more serious or fatal injuries in spring while the total amount of injuries in autumn, spring and summer seem to be pretty stable and decreased during the winter:

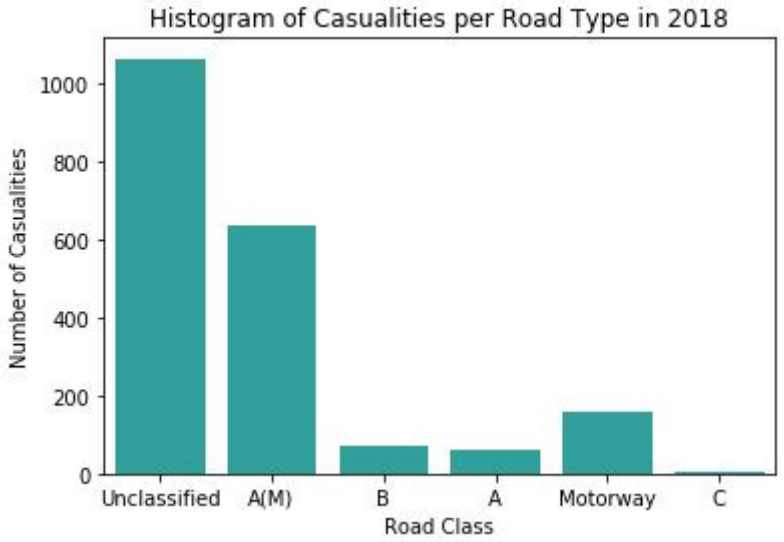


And after running the normalised value counts, I can confirm that:

Autumn	Light Injuries	0.831418
	Serious or Fatal Injuries	**0.168582**
Spring	Light Injuries	0.815385
	Serious or Fatal Injuries	**0.184615**
Summer	Light Injuries	0.833992
	Serious or Fatal Injuries	**0.166008**
Winter	Light Injuries	0.845638
	Serious or Fatal Injuries	**0.154362**

3.1.3.3 1st Road Class

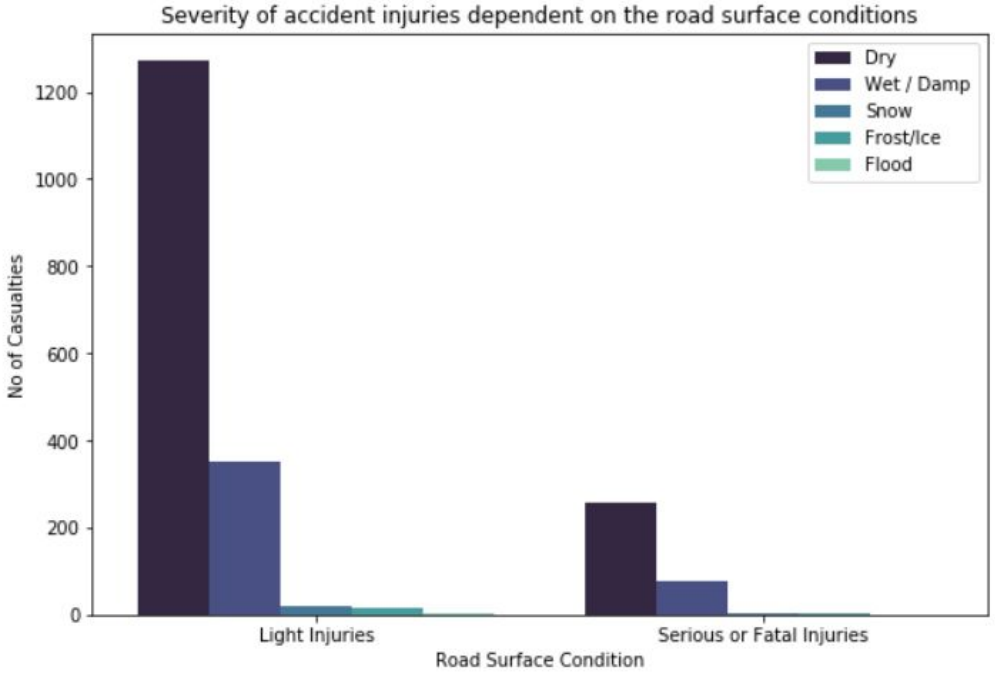
1st Road Class is the classification data for the road type reflecting the importance of the road in hierarchy: Motorway -> A(M) -> A -> B -> C.
I entered the descriptions of 1st Road Class into to data frame and plotted a simple histogram:



It clearly shows that most the rows in this column are Unclassified and therefore this column will not help us any further. Dropped the column from dataframe.

3.1.3.4 Road Surface Conditions

Quick review of the % of different surface conditions and accident rates shows that there might be correlation:



Road Surface Descriptions	Casualty Severity Description	%
Dry	Light Injuries	0.832896
	Serious or Fatal Injuries	*0.167104*
Flood	Light Injuries	1.000000
Frost/Ice	Light Injuries	0.800000
	Serious or Fatal Injuries	**0.200000**
Snow	Light Injuries	0.863636
	Serious or Fatal Injuries	0.136364
Wet / Damp	Light Injuries	0.823529

Serious or Fatal Injuries	**0.176471**

While most of the casualties occur when the road is dry, the severity of the accidents with Frost/Ice and Wet/Damp conditions is clearly higher. These might line up also with seasonality.

3.1.3.5 Lighting Conditions

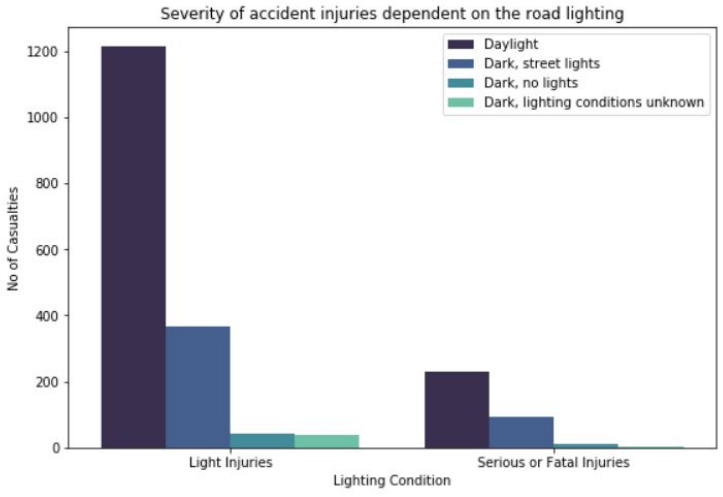
Adding in an extra descriptive column in the data frame to be able to draw easy-to-understand plots.
The description of Lighting Conditions is the following:

Lighting Conditions	1	Daylight: street lights present
Lighting Conditions	2	Daylight: no street lighting
Lighting Conditions	3	Daylight: street lighting unknown
Lighting Conditions	4	Darkness: street lights present and lit
Lighting Conditions	5	Darkness: street lights present but unlit
Lighting Conditions	6	Darkness: no street lighting
Lighting Conditions	7	Darkness: street lighting unknown

This is not very informative when considering our objective because we are concerned only with visibility and not with the condition of street lighting.
Therefore the data set has been modified to have reduced number but more relevant elements:

Lighting Conditions	1	Daylight
Lighting Conditions	2	Darkness: street lights present and lit
Lighting Conditions	3	Darkness: no street lighting
Lighting Conditions	4	Darkness: street lighting unknown

Plotted the results on bar chart:



Lighting Descriptions	Casualty Severity	%
Dark, lighting unknown	Light Injuries	0.925000
	Serious or Fatal Injuries	0.075000
Dark, no lights	Light Injuries	0.796296
	Serious or Fatal Injuries	**0.203704**
Dark, street lights	Light Injuries	0.796943
	Serious or Fatal Injuries	**0.203057**
Daylight	Light Injuries	0.840610
	Serious or Fatal Injuries	*0.159390*

From the data above we can conclude that while there are considerably more accidents in daylight, the accidents are definitely more likely to be more severe in the dark.

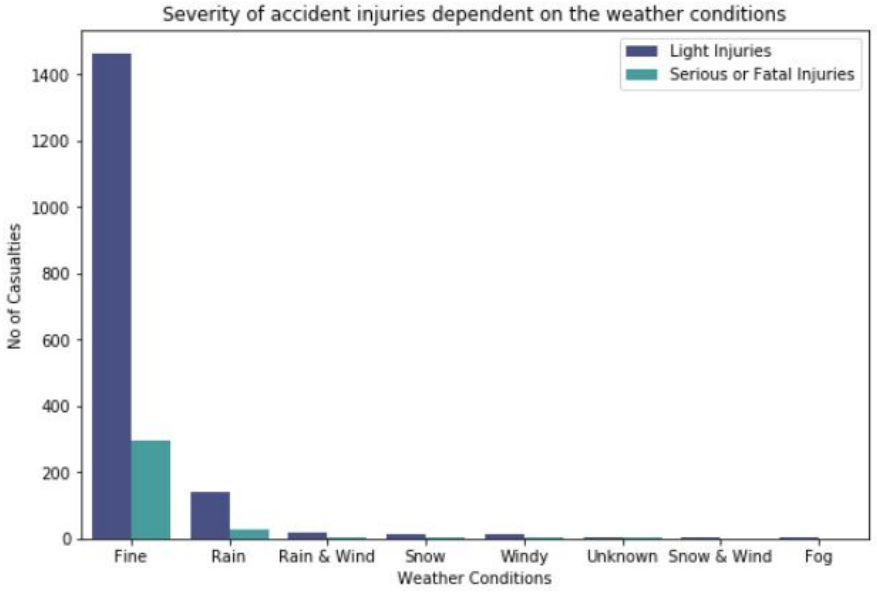
3.1.3.6 Weather Conditions

In the current data set these are as following:

Weather Conditions	1	Fine without high winds
Weather Conditions	2	Raining without high winds
Weather Conditions	3	Snowing without high winds
Weather Conditions	4	Fine with high winds
Weather Conditions	5	Raining with high winds
Weather Conditions	6	Snowing with high winds
Weather Conditions	7	Fog or mist – if hazard
Weather Conditions	8	Other
Weather Conditions	9	Unknown

I added a column to the data set called Weather Description, and merged the groups 8 and 9 together as for us there's no difference between Other and Unknown for our purposes.

Clearly most of the accidents happen with fine weather and next most common with rain:



The general division of accidents between different weather conditions:

Fine	0.880702
Rain	0.083208
Rain & Wind	0.010025
Snow	0.008020
Windy	0.008020
Unknown	0.006015
Fog	0.002005
Snow & Wind	0.002005

Foggy or stormy weather seem to cause more severe accidents:

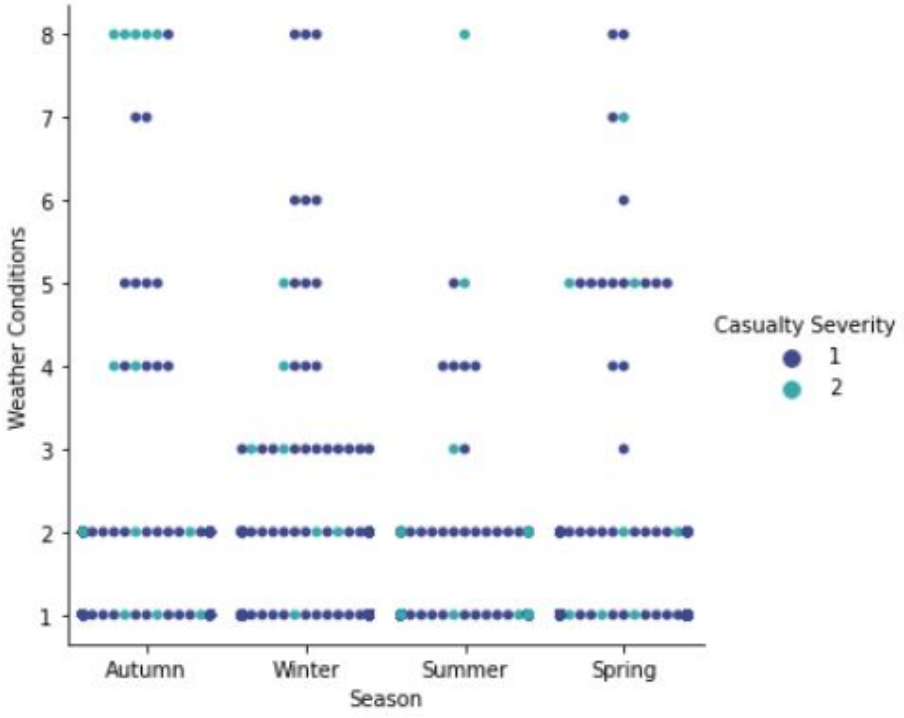
Weather Descriptions	**Casualty Severity**	**%**
Fine	*Light Injuries*	*0.832100*
	Serious or Fatal Injuries	*0.167900*
Fog	Light Injuries	0.750000
	Serious or Fatal Injuries	**0.250000**
Rain	Light Injuries	0.849398
	Serious or Fatal Injuries	0.150602
Rain & Wind	Light Injuries	0.800000
	Serious or Fatal Injuries	**0.200000**
Snow	Light Injuries	0.812500
	Serious or Fatal Injuries	0.187500
Snow & Wind	Light Injuries	1.000000
Unknown	Light Injuries	0.500000
	Serious or Fatal Injuries	0.500000
Windy	Light Injuries	0.812500
	Serious or Fatal Injuries	0.187500

All in all the accident rate nor severity seem to be affected too much about weather. The rain and dry weather numbers look pretty similar, and there are not enough snowy-foggy-windy days to draw clear conclusions.

I used the weather conditions together with Time of day to see if there's maybe correlation, but the outcome was negative. The weather didn't affect the severity or number of accidents during the different time of days.

However, there seems to be a connection between the weather condition and season:

- The higher severity value seems to be on top left hand corner of the chart, but as 8 is Unknown value, we cannot use it in prediction
- But there is clearly more accidents in Winter with weather condition 3 (= Snow) and in Spring with weather condition 5 (= Rain and Wind)



3.1.4 Casualty's Personal Features

A feature by feature overview of factors that related to the person of casualty.

3.1.4.1 Type of Vehicle

It is only fair to assume that some vehicles are more dangerous to travel in than others. When we are in a big vehicle, secured with a seat belt and surrounded by lots of room we are less likely to have serious impact injuries than when we are on a motorbike without any serious protection.

The original list of vehicle descriptions is long and complicated:

Category	Table Value	Description
Type of Vehicle	1	Pedal cycle
Type of Vehicle	2	M/cycle 50cc and under
Type of Vehicle	3	Motorcycle over 50cc and up to 125cc

Type of Vehicle	4	Motorcycle over 125cc and up to 500cc
Type of Vehicle	5	Motorcycle over 500cc
Type of Vehicle	6	[Not used]
Type of Vehicle	7	[Not used]
Type of Vehicle	8	Taxi/Private hire car
Type of Vehicle	9	Car
Type of Vehicle	10	Minibus (8 – 16 passenger seats)
Type of Vehicle	11	Bus or coach (17 or more passenger seats)
Type of Vehicle	12	[Not used]
Type of Vehicle	13	[Not used]
Type of Vehicle	14	Other motor vehicle
Type of Vehicle	15	Other non-motor vehicle
Type of Vehicle	16	Ridden horse
Type of Vehicle	17	Agricultural vehicle (includes diggers etc.)
Type of Vehicle	18	Tram / Light rail
Type of Vehicle	19	Goods vehicle 3.5 tonnes mgw and under
Type of Vehicle	20	Goods vehicle over 3.5 tonnes and under 7.5 tonnes mgw
Type of Vehicle	21	Goods vehicle 7.5 tonnes mgw and over
Type of Vehicle	22	Mobility Scooter
Type of Vehicle	90	Other Vehicle
Type of Vehicle	97	Motorcycle - Unknown CC

Have simplified the data as well as added a new column to the data set:

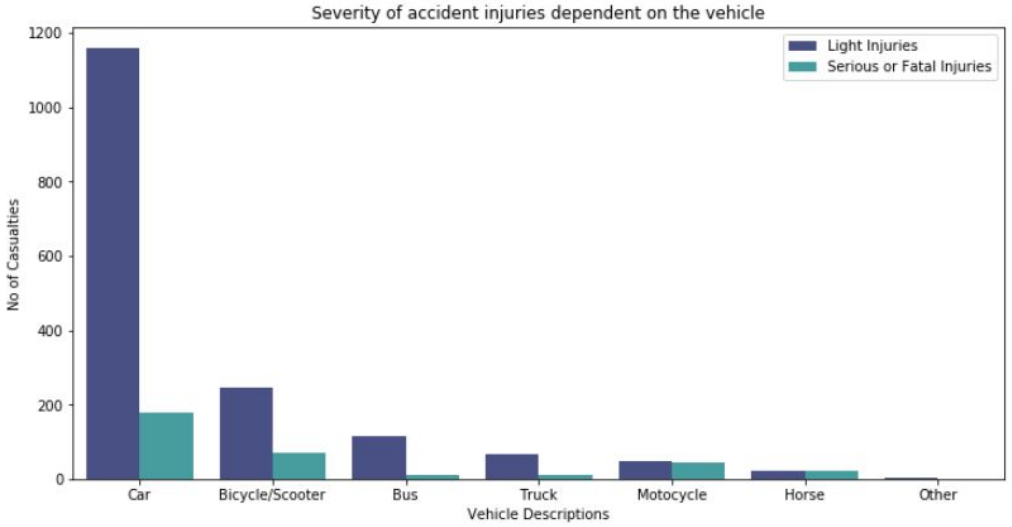
Category	Table Value	New Description / Column Value
Type of Vehicle	1	Bicycle / Scooter
Type of Vehicle	2	Motorcycle
Type of Vehicle	3	Car / Minibus
Type of Vehicle	4	Bus
Type of Vehicle	5	Horse
Type of Vehicle	6	Big & Slow (Agricultural vehicles)
Type of Vehicle	7	Tram / Light Rail
Type of Vehicle	8	Truck
Type of Vehicle	9	Other

The casualty count for each vehicle type:

Car	1337
Bicycle/Scooter	314
Bus	127
Motorcycle	93
Truck	80

Horse	40
Other	4

The division of Light vs Serious/Fatal Injuries per vehicle type that shows it's relatively high potential to end up in the accident when you are on bicycle or scooter, and on horse:



The above graph is displayed in percentages below which shows clearly that while *riding a bicycle or scooter* is more likely to cause more severe casualties compared to a car or bus, it is **the horse and motorcycle** that are the most dangerous options for travelling.

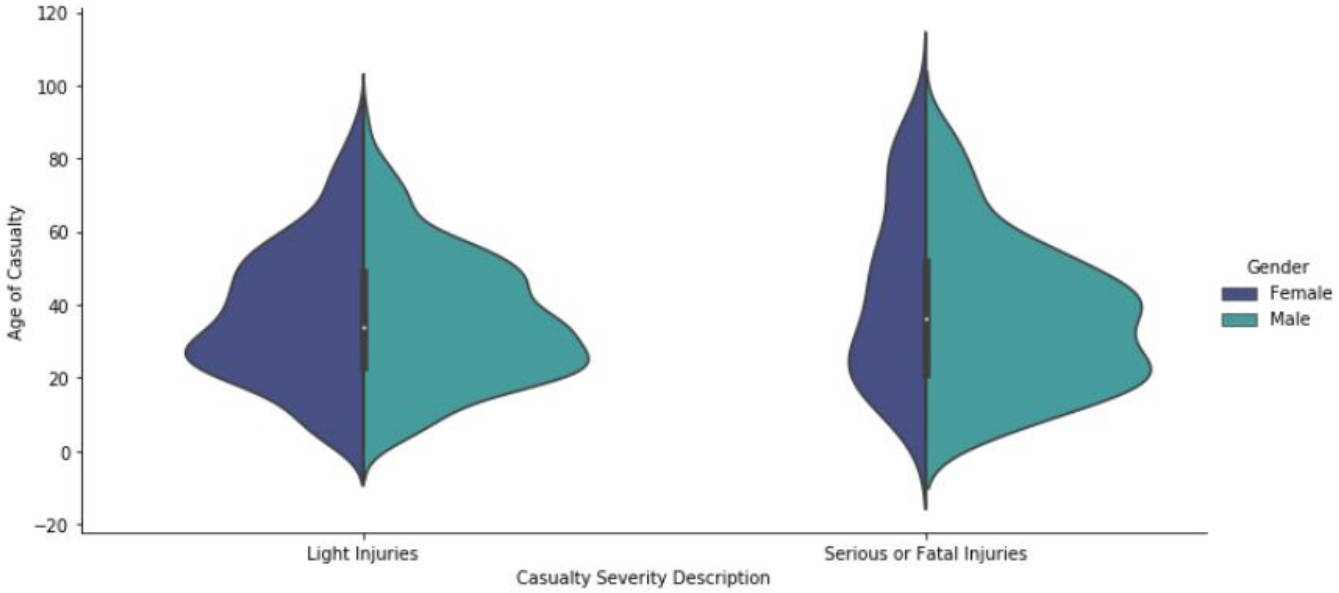
Vehicle Descriptions	**Casualty Severity Description**	**%**
Bicycle/Scooter	*Light Injuries*	*0.780255*
	Serious or Fatal Injuries	*0.219745*
Bus	Light Injuries	0.905512
	Serious or Fatal Injuries	0.094488
Car	Light Injuries	0.866118
	Serious or Fatal Injuries	0.133882
Horse	**Light Injuries**	**0.500000**
	Serious or Fatal Injuries	**0.500000**
Motorcycle	**Light Injuries**	**0.526882**
	Serious or Fatal Injuries	**0.473118**
Other	Light Injuries	0.750000
	Serious or Fatal Injuries	0.250000
Truck	Light Injuries	0.850000
	Serious or Fatal Injuries	0.150000

3.1.4.2 Age and Gender of Casualties

Added in a descriptive data column for gender and computed the counts:

Male	1176
Female	819

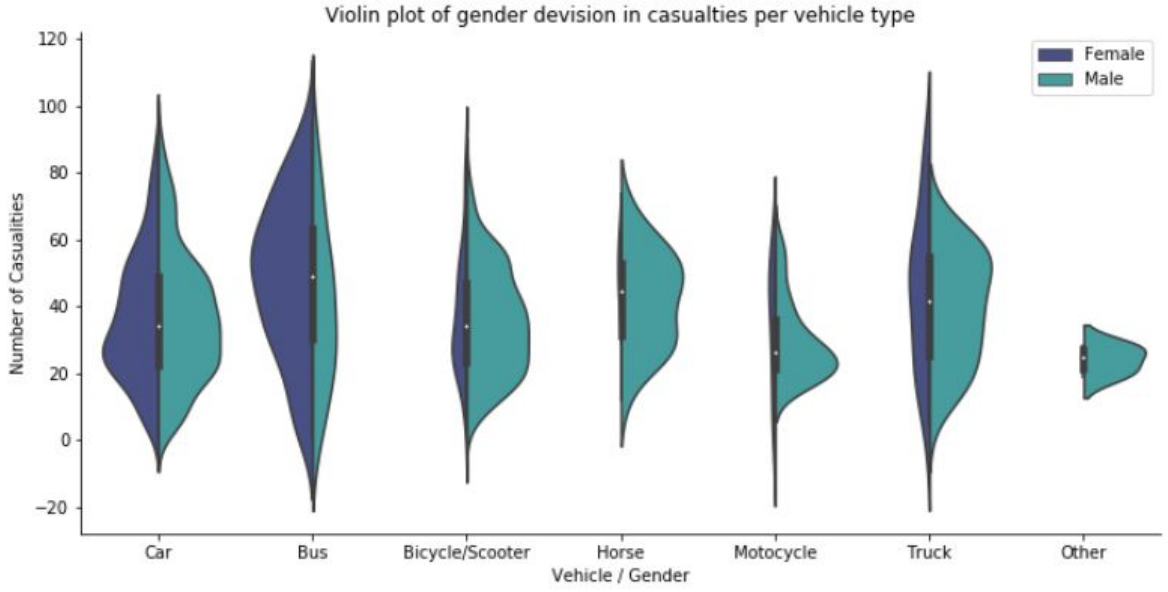
Then created a plot seeing how is the casualty severity divided between genders through age groups:



Men are more likely being casualties of accidents, but they are two times more likely to suffer serious or fatal injuries:

Gender	**Casualty Severity Description**	**%**
Female	Light Injuries	0.894994
	Serious or Fatal Injuries	**0.105006**
Male	Light Injuries	0.786565
	Serious or Fatal Injuries	**0.213435**

There are also clear correlations between gender, age and accidents with specific vehicles like motorcycles, trucks, bicycles etc.



3.1.5 Final Data Set

Final Data Set has 1995 rows and 25 following columns:
'Accident Ref No',
'Number of Vehicles', 'Number of Casualties'
'Casualty Severity', 'Casualty Severity Description', 'CS Value',
'SS Value', 'Severity Level',
'Accident Date', 'Season', 'Season Code',
'Time (24hr)', 'Time of day', 'Time of Day Values'
'Road Surface', 'Road Surface Description',
'Lighting Conditions', 'Lighting Description',
'Weather Conditions', 'Weather Description',
'Type of Vehicle', 'Vehicle Description',
'Sex of Casualty', 'Gender',
'Age of Casualty'

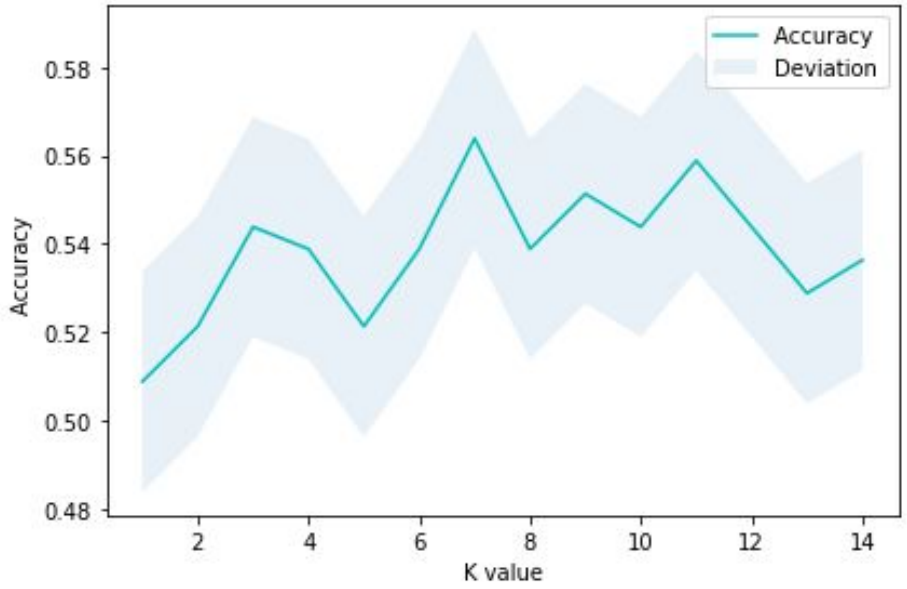
3.2 Data Modelling

As we are trying to predict a category, we need a ML classification model. We also have more than two categories, so ideally we would use a model that is not the best only for binary categorisation (e.g. Logistic Regression). That leaves us with KNN and Decision Tree models.

For the sake of certainty, both models were plotted and accuracy calculated. The test-train split for the in-sample data testing was set to 80-20 because the dataset is relatively small, so more data for training, the better.

3.2.1 K-Nearest Neighbor KNN

The KNN model worked, but with relatively low accuracy. Several test-train splits were tested (70-30, 75-25, 80-20), but 80-20 split still produced the best accuracy which was 56.4% with k=7



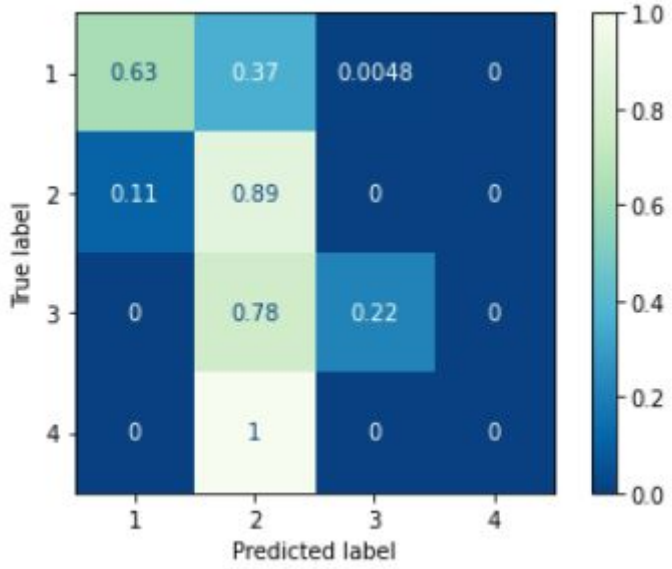
KNN Accuracy: 0.5639097744360902
KNN Jaccard index: 0.56
KNN F1-score: 0.55

3.2.2 Decision Tree

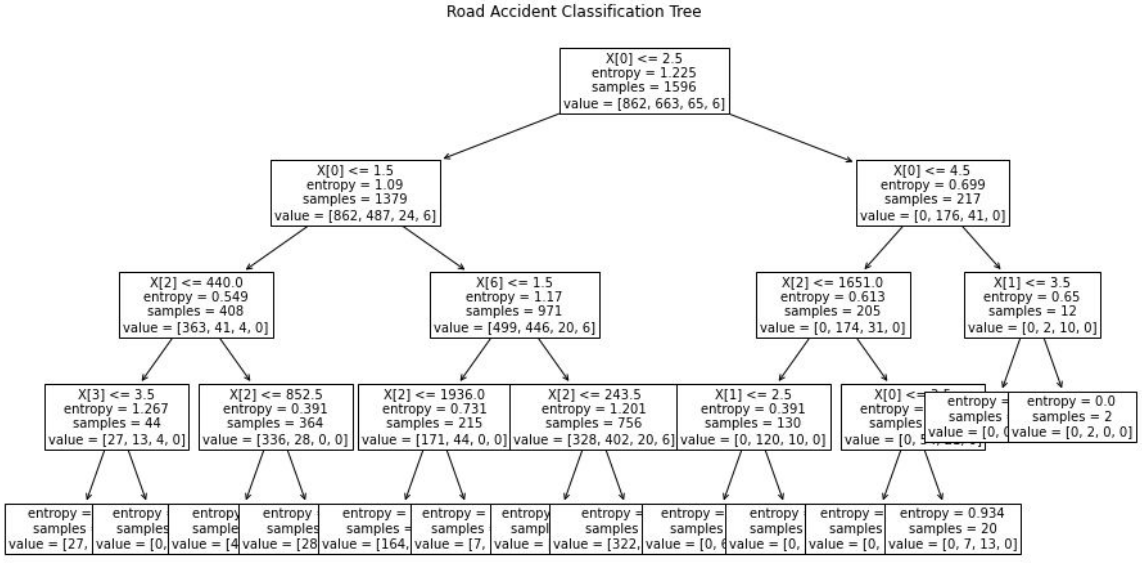
Decision Tree model worked considerably better. The 80-20 training-test split produced:
Tree Accuracy Score: 0.74
Tree Jaccard index: 0.74
Tree F1-score: 0.73

The confusion matrix with following results:

- the model classifies 37% Level 1 cases as Level 2 i.e. it assesses that the accident is more severe than it actually is
- the model classifies 11% Level 2 cases as Level 1, i.e. it assesses that the accident is less severe than it actually is
- the model mis-classifies the Level 3 and Level 4 accidents mostly as Level 2, which is explained by the shortage of relevant data. Because the amount of accidents with Level 3 and Level 4 is very small, the model treats them as outliers and naturally classifies the accident in the next most likely group, which is Level 2.



The classification tree itself:



4 Model Evaluation & Results

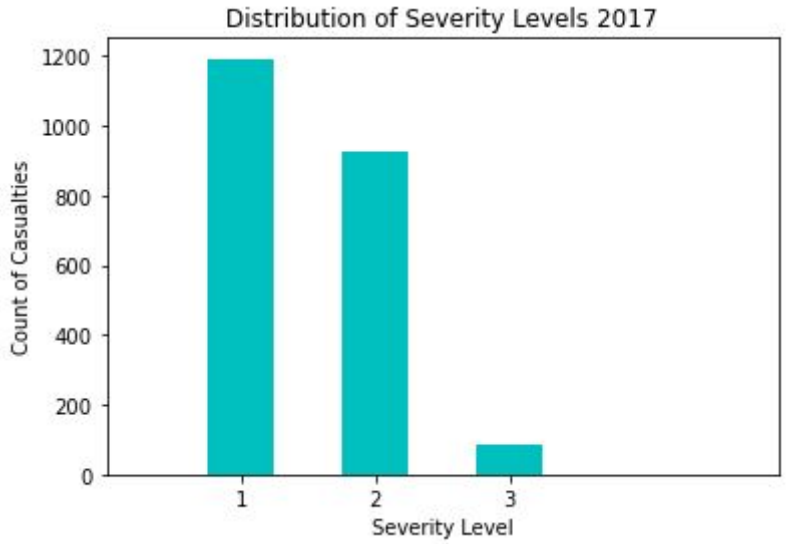
Model evaluation is done with out-of-sample data. The data used is from the same UK Government data source, but for year 2017. This assures the continuity of data as it was collected in the same area under the same circumstances.

The original 2017 data set was not formatted into numerical values, and only contained feature descriptions as shown on the screenshot below. Therefore the whole data set had to be processed feature by feature in order to match the model's df_2018 format.

Original data:

	Accident Ref No	Number of Vehicles	Accident Date	Time (24hr)	Road Surface	Lighting Conditions	Weather Conditions	Type of Vehicle	Casualty Severity Description	Sex of Casualty	Age of Casualty
0	3AP0313	1	3/17/2017	815	Dry	Daylight: Street lights present	Other	Car	Serious	Female	61
1	3BE0850	2	1/14/2017	1330	Dry	Daylight: Street lights present	Fine without high winds	Pedal cycle	Slight	Male	36
2	4110858	2	1/1/2017	805	Wet/Damp	Daylight: Street lights present	Fine without high winds	Car	Slight	Male	32
3	4110858	2	1/1/2017	805	Wet/Damp	Daylight: Street lights present	Fine without high winds	Car	Slight	Male	30
4	4111495	2	1/1/2017	1705	Wet/Damp	Darkness: Street lights present and lit	Raining without high winds	Car	Slight	Female	26

For the classification the Severity Levels are used, and in the 2017 test data these are distributed as following:



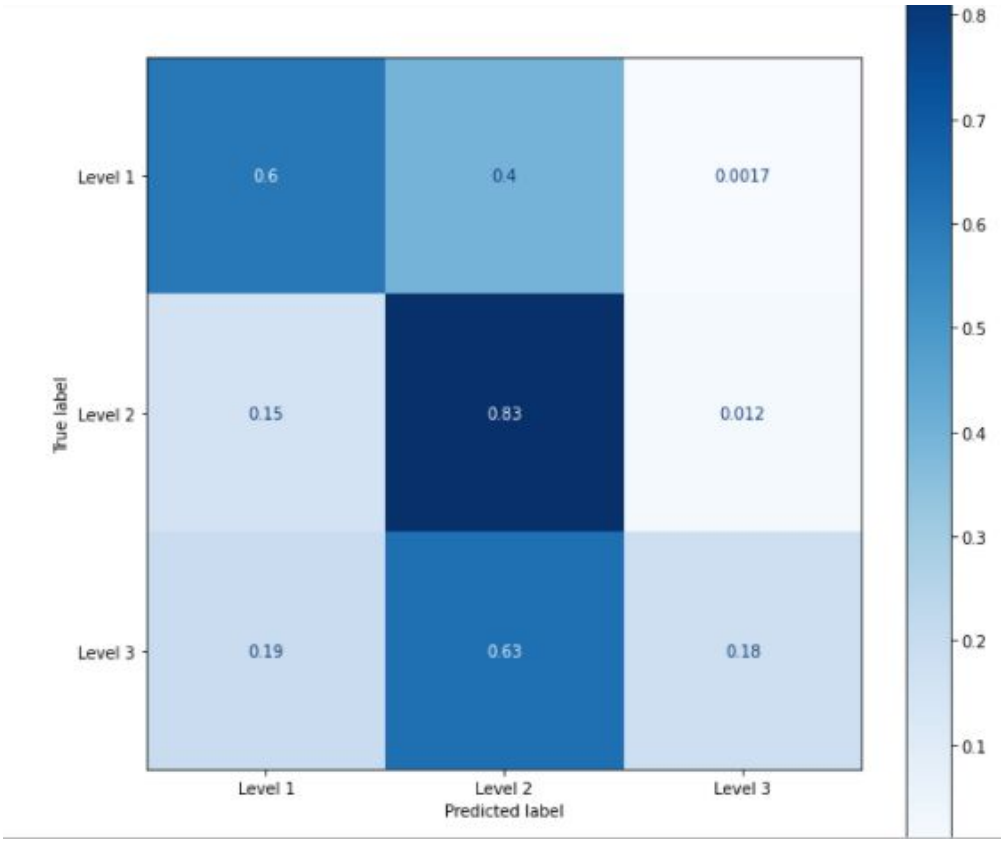
4.1 Results

The Tree model was tested with the out-of-sample data with the following results:

F1 score = 0.678
Jaccard Score = 0.683

Both of these values are lower than the testing results with in-the-sample data set (values 0.74 and 0.73 respectively).

Confusion Matrix for model prediction vs 2017 real data:
- the model classifies 40% Level 1 cases as Level 2 i.e. it assesses that the accident is more severe than it actually is
- the model classifies 15% Level 2 cases as Level 1, i.e. it assesses that the accident is less severe than it actually is
- the model mis-classifies the Level 3 accidents mostly as Level 2 and even Level 1, which again is explained by the shortage of relevant data.



5 Discussion & Conclusion

In this project it was analysed if the accident severity could be predicted considering environmental features at accident time, and personal features of casualties.

While the machine learning model created with Decision Tree algorithm is relatively good for the initial results, it has room for improvement.
We can see that the amount of data is not sufficient and because the Level 3 and Level 4 accidents are treated as outliers, it creates an imbalance in the model. For example the 2017 testing data set did not have a single Level 4 accident.
In order to make the model better, all the available data since 2009 from the Leeds could be formatted and used to train the model. That would make sure the data set is bigger and therefore more accurate. Another option would be adding the data from other similar towns in UK in order to increase the set.

Also the data format might be a limiting factor for more accurate prediction: there are limited amount of analysis one can perform with categorical data. For example it might be worth exploring additional data sets with weather data and merge it with current data set in order to get the Weather, Road and Lighting Conditions as continuous numbers (e.g. Rainfall amount 0 - 50mm, instead of categories Fair / Rain)

The current model could be used by emergency services with caution. However, the results can be used only in similar conditions i.e. in the northern hemisphere in similar climates because it takes into account for example seasonality and weather conditions like snow. For any other part of the world, the model would need to be re-trained with relevant data.

The conclusion is definitely positive, and the severity prediction can be done if the data set is sufficient. Features like time, date, weather, lighting and road conditions, as well as number of vehicles and people involved, drivers' age, gender, and the choice of vehicle can be used to predict the Accident Severity with reasonable accuracy using ML models.