Problem Description

Traffic accidents are a common occurrence of everyday life. These can cause material losses, personal injury, emotional distress, traffic disruption and unfortunately in some cases death. In many countries and cities there is an increased amount of cars on the road. Various reasons can be attributed to this rise in traffic:

- Increased population
- Lack or unreliable public transport
- Low cost of cars
- Increased disposable incomes

Due to the increase of cars on the road and potential accidents, it is essential to understand the causes of accident severity which would be useful for various bodies such as police departments, hospitals, insurance companies, transport companies, among many others.

These stakeholders will benefit of these predictions by being able utilise their resources more efficiently. For example hospitals would be able to have staff available only in times when conditions are higher for more severe accidents and reduce staff when it is lower. This would be a similar situation for police departments.

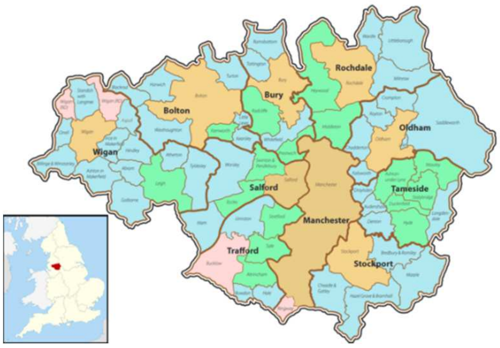
Insurance companies would be able to understand when and why high severity accident happens. For example if road condition is a large factor, they might incentivise customers to have a more stringent tyre changes or use snow tyres.

Transport companies like taxis might avoid certain days or conditions when a high number of severe accidents happen.

The objective of the capstone project is to predict the severity of a traffic accident in Greater Manchester, England. Using data science and machine learning techniques, this project will analyse accident data from 2018 to understand the factors that affect the severity of an accident.

Data Description and Approach

The data being used has been gathered by the UK Department of Transport and it includes all traffic accidents during 2018 in the Greater Manchester area. Greater Manchester is a large metropolitan area in the north of England with an approximate population of 2.8 million and includes 10 boroughs: Bolton, Bury, Oldham, Rochdale, Stockport, Tameside, Trafford, Wigan and the cities of Manchester and Salford.



The dataset is publicly available on the following site: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

The data includes more than 122,000 data points and more than 30 attributes such as:

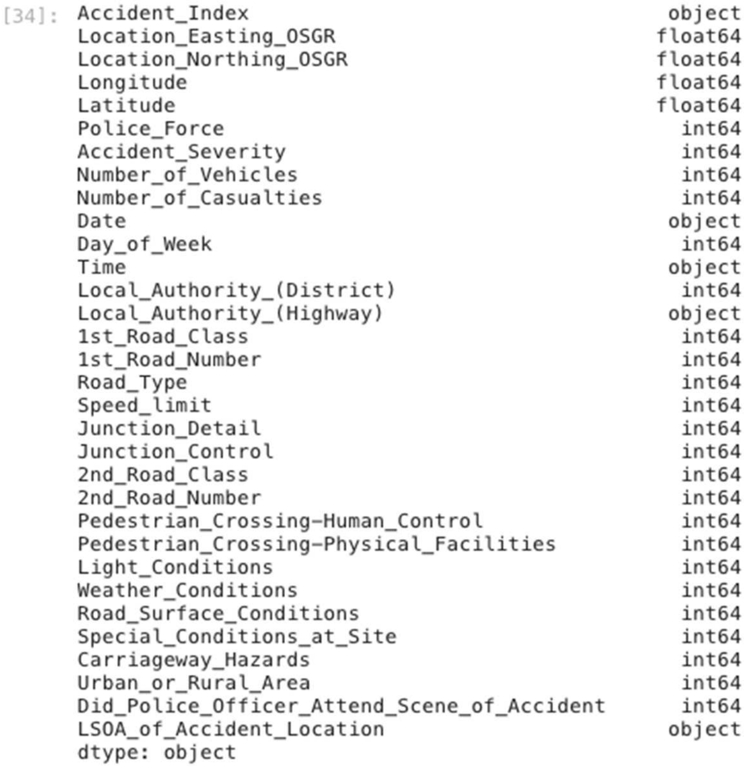
- Accident data: location, number of vehicles involved, date
- Environmental data: light conditions, weather conditions, road conditions,
- Others: local authority, special conditions, police attendance

The dataset uses 3 different attributes to identify the severity of the accidents – Fatal (1), Serious (2) and Slight (3).

To solve the problem, firstly the plan is to understand the available data and the available attributes. Next, look for trends on which attributes are the most relevant to use. Afterwards, ensure the data is ready for modelling including data balancing, filling missing data and generally cleaning the dataset. Afterwards I will review which variables have a strong correlation to the one we want to predict using correlation analysis. Since the aim of the analysis is to predict and categorise accident severity, I will use a model using supervised machine learning techniques. Lastly, evaluate the model to ensure the business objectives are achieved using metrics such as recall, precision, and F1-score.

Methodology

As mentioned previously, the data is all the traffic accidents in Greater Manchester for 2018. The information was gathered by the UK Department of transport. It includes 32 attributes and 122,635 rows. In order to start doing the analysis, firstly it needed to be cleaned to ensure all values are populated and it includes factors that might affect the severity of the accident.



```
[34]: Accident_Index                                    object
      Location_Easting_OSGR                            float64
      Location_Northing_OSGR                           float64
      Longitude                                        float64
      Latitude                                         float64
      Police_Force                                       int64
      Accident_Severity                                  int64
      Number_of_Vehicles                                 int64
      Number_of_Casualties                               int64
      Date                                              object
      Day_of_Week                                        int64
      Time                                              object
      Local_Authority_(District)                         int64
      Local_Authority_(Highway)                         object
      1st_Road_Class                                     int64
      1st_Road_Number                                    int64
      Road_Type                                          int64
      Speed_limit                                        int64
      Junction_Detail                                    int64
      Junction_Control                                   int64
      2nd_Road_Class                                     int64
      2nd_Road_Number                                    int64
      Pedestrian_Crossing-Human_Control                  int64
      Pedestrian_Crossing-Physical_Facilities            int64
      Light_Conditions                                   int64
      Weather_Conditions                                 int64
      Road_Surface_Conditions                            int64
      Special_Conditions_at_Site                         int64
      Carriageway_Hazards                                int64
      Urban_or_Rural_Area                                int64
      Did_Police_Officer_Attend_Scene_of_Accident        int64
      LSOA_of_Accident_Location                         object
      dtype: object
```

Considering all these characteristics, some needed to be removed for the following reasons:

- Attributes related to the police or authority: Police_Force, Local_Authority_(District), Did_Police_Officer_Attend_Scene_of_Accident
- Attribute too specific to be relevant to the study: Special_Conditions_at_Site, Pedestrian_Crossing-Human_Control, Location_Easting_OSGR, Location_Northing_OSGR, 1st_Road_Class, 1st_Road_Number

After this clean-up, the study ended up with 6 relevant attributes and the predicted one:

- Day of Week: Monday through Sunday
- Road Type: One way, single carriageway, double carriageway, slip road or roundabout
- Light Conditions: Daylight, dark-lit and dark-unlit
- Urban or Rural Area
- Road Surface Conditions: Dry, Wet, Snow, Ice or Flood
- Weather Conditions: Fine, Rain, Snow or Fog
- **Accident Severity: Fatal, Serious or Slight**

Another clean-up that had to be done was to remove any unknown, missing or other data. In some cases it needed to be found through specific number in the data. For example in the case of Road Condition, the data was gathered as follows:

code	label
1	Dry
2	Wet or damp
3	Snow
4	Frost or ice
5	Flood over 3cm. deep
6	Oil or diesel
7	Mud
-1	Data missing or out of range

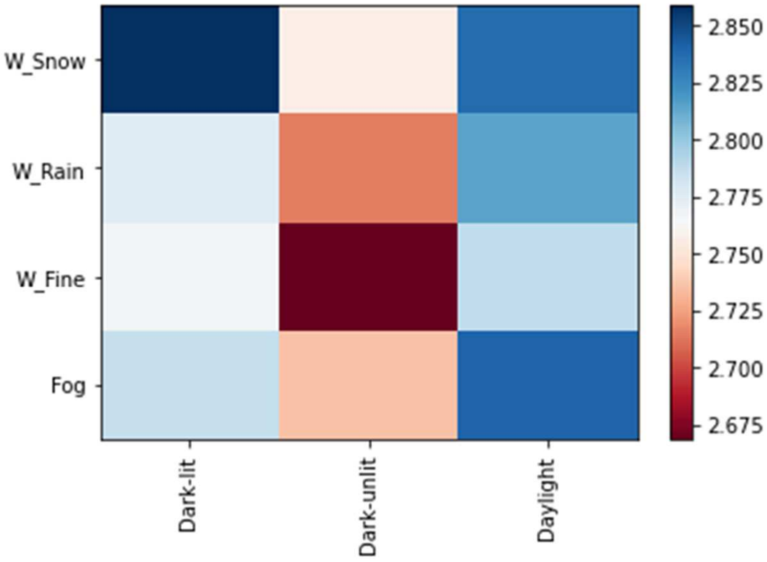
In this case, I needed to remove all lines with a "-1" value.

It was decided to remove any data that was missing since it did not make much of a difference to the total number of data points. The original number was 122,635 and after the clean-up it became 116,806 so less than 5% data loss.

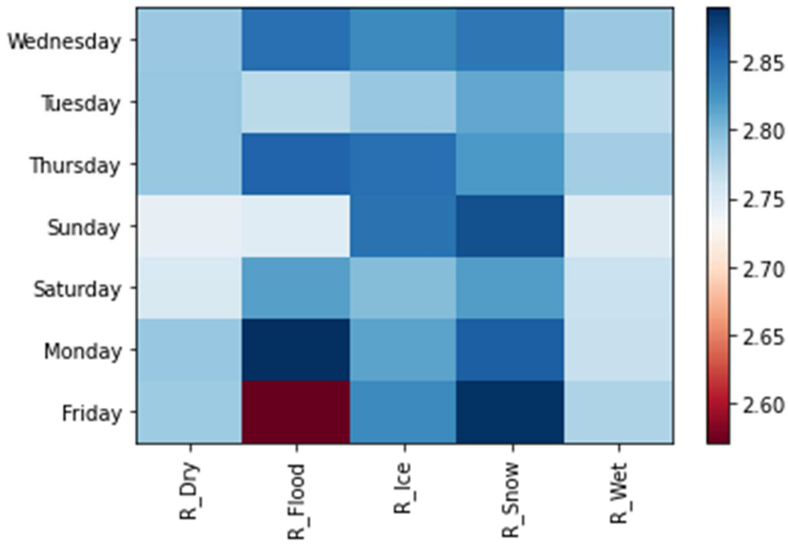
Once the data was corrected, I started looking for relationships of the different variables into the severity of the accident. Fatal accidents are very uncommon, accounting for less than 15% of all accidents in the area. Looking at the different heatmaps I was able to find some correlations. Severity of accidents go from

- 3 – Slight
- 2 – Serious
- 1 – Fatal.

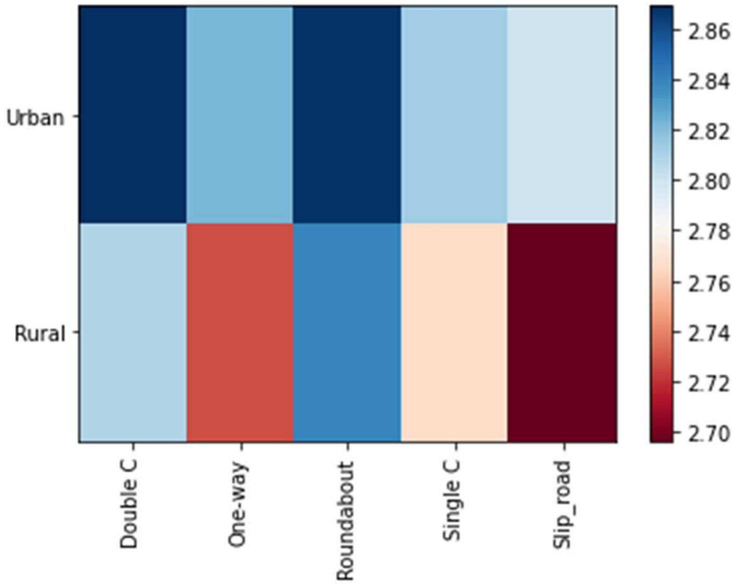
Below we are comparing weather and lighting conditions. It is very noticeable that dark-unlit areas are the most common for severe accidents, while daylight has the less severe accidents. Weather seems to not be a very relevant variable since it has a diverse range of values.



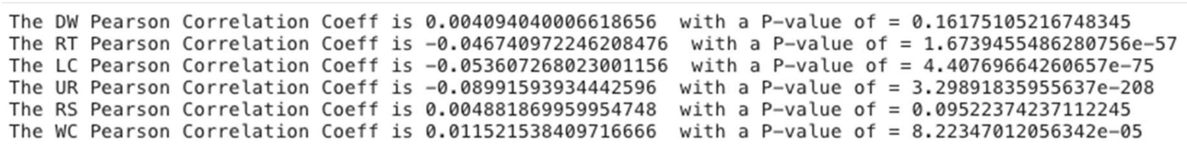
The heatmap below relates the day of the week and the road condition. It seems the days of the week are not very relevant. Originally I was expecting a trend of more severe accidents during the weekend, but that does not seem to be the case. Regarding the road condition, it seems the most severe accidents happen in wet and dry conditions. It is surprising that some happen on dry conditions, maybe caused by increase in speeds and less caution taken by the drivers.



The heatmap below relates the location being rural or urban and the road type. It seems than rural roads are more dangerous, since in all cases it has a higher number than its urban equivalent. The type of road seems to have lots of relevance, double carriageways and roundabouts are the safest and single carriageways, slip roads and one way streets being responsible for more severe accidents.



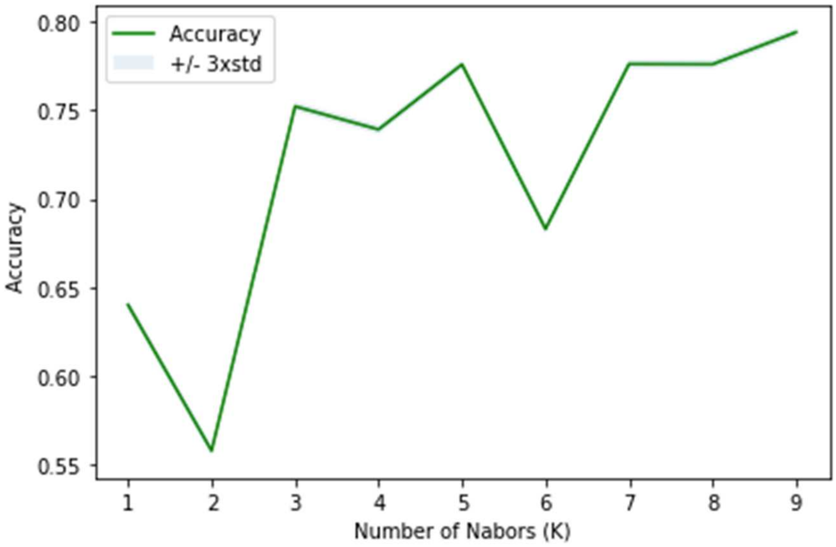
In total, there are very low correlations between each of the variables and the severity of the accident. However looking at the Pearson Correlation Equivalent, it is clear that two of them are negligible: Day of Week (DW) and Road Surface Condition (RS).



```
The DW Pearson Correlation Coeff is 0.004094040006618656  with a P-value of = 0.16175105216748345
The RT Pearson Correlation Coeff is -0.046740972246208476  with a P-value of = 1.6739455486280756e-57
The LC Pearson Correlation Coeff is -0.053607268023001156  with a P-value of = 4.40769664260657e-75
The UR Pearson Correlation Coeff is -0.08991593934442596  with a P-value of = 3.29891835955637e-208
The RS Pearson Correlation Coeff is 0.004881869959954748  with a P-value of = 0.09522374237112245
The WC Pearson Correlation Coeff is 0.011521538409716666  with a P-value of = 8.22347012056342e-05
```

The prediction was made with 4 different types of modelling: K-Nearest Neighbour, Decision Tree, Logistic Regression and Support Vector Machines. The data was split into a test and train, with test being 20% for all models.

With KNN I discovered the best number for prediction was k=9, which gave me a prediction of 0.79 for both train and test data.



Likewise a similar process was followed to model the severity of accidents using Decision Tree, Logistic Regression and Support Vector Machines, which gave a very similar prediction results. Likewise the provided a weighted F1 prediction of 0.702.

Results

In general, we have discovered that the main culprits into the severity of an accident in Greater Manchester are:

- Road Type: One way, single carriageway, double carriageway, slip road or roundabout
- Light Conditions: Daylight, dark-lit and dark-unlit
- Urban or Rural Area
- Weather Conditions: Fine, Rain, Snow or Fog

Using these variables, we can get an 80% accuracy model of the severity of the accident using K-Nearest Neighbour, Decision Tree, Logistic Regression and Support Vector Machines.

Discussion

Based on these results, a few recommendations can be made:

- Accidents in unlit darkness are more serious than in lit darkness. Further analysis can be made to see where these accidents are more common to possibly invest in public lighting.
- Accidents are more severe in winter due to the darkness, so maybe temporary staff might be needed in health departments.
- Accidents are more severe in wet conditions, so might be relevant for insurance companies to incentivise the purchase and use of wet tread tyres. Another option would be to review the speed limits in some of these zones.
- Police and hospitals might try to review weather conditions to assess the number of personnel they might need on-duty.
- Accidents are also more severe in slip roads so maybe design engineers would be interested in designing them with better visibility or lower speed limits.

This report and analysis would be interesting for various stakeholders such as insurance companies, hospitals, police departments, road designers, government councils, etc. The prediction using the above mentioned variables would give us an accurate estimate of the severity of the accident.

Conclusion

In conclusion, the Capstone allowed me to go through the process of identifying a problem, accessing publicly available data and have a good understanding of it. Afterwards cleaning and preparing it for analysis and looking for relations between the variables which can give us a good prediction. Lastly looked at different modelling for prediction including K-Nearest Neighbour, Decision Tree, Logistic Regression and Support Vector Machines. Stakeholders can look at these variables and be better prepared to encounter severe accidents in Greater Manchester and act to reduce the severity.