ROAD SAFETY AND SEVERITY ANALYSIS
IN GREAT BRITAIN

Miguel Rivera

September 12, 2020

1. INTRODUCTION

1.1. Backgrounds

According to the Department for Transport of Great Britain, there were 1.671 reported road deaths in 2018, 23.165 serious injuries in road traffic accidents reported to the police and 97.799 slight injuries. All this adds up to 122.635 reported road accidents, this is 6% lower than in 2017 and is the lowest level on record. There is no single underlying factor that drives road accidents. Instead, there are several influences. These include:

- The distance people travel.

- The mix of transport modes used.

- Behavior of drivers, riders, and pedestrians.

- The mix of groups of people using the road (e.g. changes in the number of newly qualified or older drivers)

- External effects such as the weather, which can influence behavior (for instance, encouraging / discouraging travel, or closing roads) or change in the risk on roads (by making the road surface more slippery).

The patterns involved in dangerous crashes and road accidents can be helpful in developing road safety policies or greater awareness of the driver himself. The study of these patterns, causes and severity of injuries can predict the severity or probability of an accident, which would be extremely useful for the well-being of any driver or pedestrian

1.2. Business Understanding

The analysis of these accidents causes and severity of injuries can be used to predict the severity or probability of an accident. Therefore, it would be very useful for the well-being of any driver or pedestrian, to offer the possibility that, knowing the risks, the probability of an accident, the most dangerous routes and the less favorable weather conditions, among many other factors, people drive carefully or even change their trip.

1.3. Interest

The main objective is to identify the severity or probability of a possible accident, so this work can be useful for anyone interested in assessing risks while driving, the traffic control departments or simply for anyone who drives a vehicle, offering the possibility that, knowing the risks and probabilities of an accident, people drive with more care. or even change their trip. Undoubtedly, this would help all drivers to have a better perspective of possible accidents, the most dangerous routes, the less favorable weather conditions and with all these data, to make a more informed decision.

2. DATA

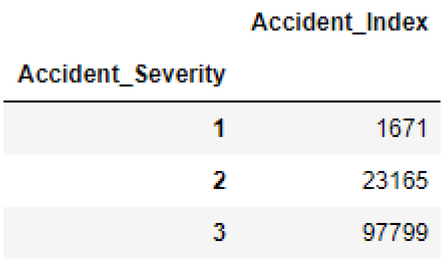
2.1. Data Source.

This project is developed over the data provided by the UK open data website, section road safety data (https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data) for year 2018: **ROAD SAFETY DATA - ACCIDENTS 2018** and his Definition Dictionary **STATS19 Variable Lookup data guide**

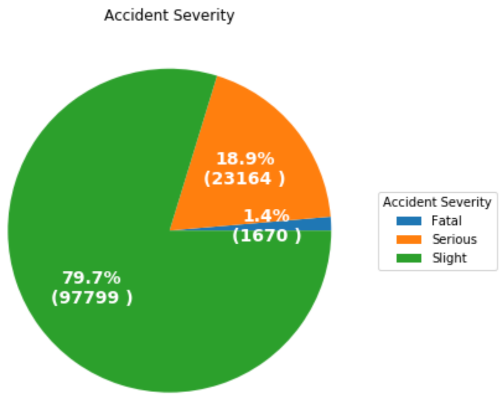
The downloaded data extracted from a single source allows a better management of the information. This repository has the advantage of having a structured and well-formatted data set; however, it presents labels with missing data encoded as -1, which are not easily identifiable when searching for null or blank records. The set presents a good number of variables to build a model, some more predictive than others and although it presents useful geospatial data to produce visualizations, I will see if they are useful for the present project.

Our dataset consists of 122635 files. Each row represents an accident, and for each one is documented 32 specific features involved in it. We know that our main metric will be **Accident_Severity**, so checked their information in our data frame:

No. Accidents by Severity



	Accident_Index
Accident_Severity	
1	1671
2	23165
3	97799

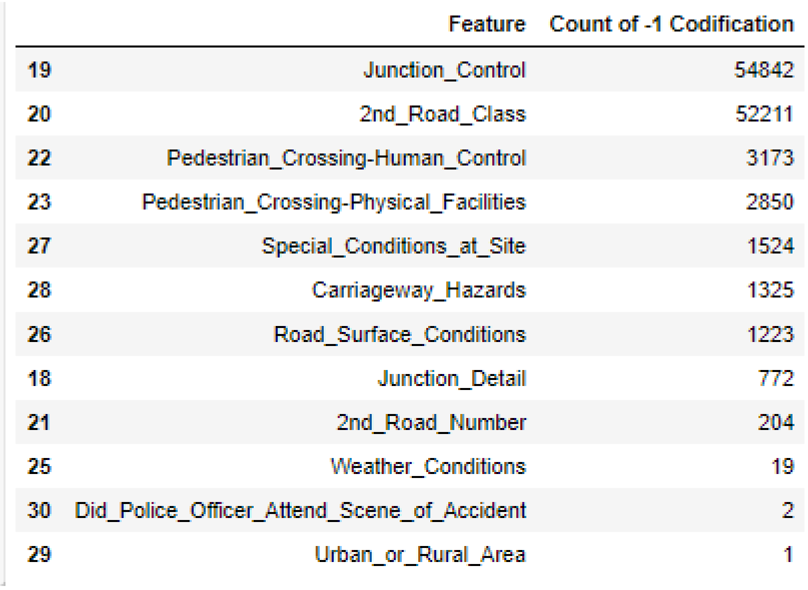


As expected, we have far fewer accidents category 1 (**1,4 %**), which are the fatalities. Category 2 (**18,9 %**) are the severe accidents, and category 3 (**79,7 %**) are slight injuries and none missing values into the main metric.

2.2. Data Cleaning.

Downloaded data extracted from a single source enables better information management. This repository has the advantage of having a well-formatted and structured data set; however, it presents labels with missing data encoded as -1, which are not easily identifiable when searching for null or blank records. The set presents a good number of variables to build a model, some more predictive than others and although it presents useful geospatial data to produce visualizations, they are not useful for the present project.

Once the file is imported, the first thing to do is calculate and evaluate the impact of these missing values coded as -1 and if the associated characteristics are necessary or not in the construction of the model, in the initial calculation we have the following quantities per characteristic :



	Feature	Count of -1 Codification
19	Junction_Control	54842
20	2nd_Road_Class	52211
22	Pedestrian_Crossing-Human_Control	3173
23	Pedestrian_Crossing-Physical_Facilities	2850
27	Special_Conditions_at_Site	1524
28	Carriageway_Hazards	1325
26	Road_Surface_Conditions	1223
18	Junction_Detail	772
21	2nd_Road_Number	204
25	Weather_Conditions	19
30	Did_Police_Officer_Attend_Scene_of_Accident	2
29	Urban_or_Rural_Area	1

To deal with, its necessary to perform some transformations on the dataset looking for eliminate the impact of this condition:

Junction Control: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to "Not at junction or within 20 meters"

2nd Road Class: It is not of interest in this analysis, so the field is removed

Pedestrian Cross – Human Control: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to " None within 50 meters "

Pedestrian Cross – Physical Facilities: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to " No physical crossing facilities within 50 meters"

Special Conditions at site: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to "None"

Carriageway Hazards: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to "None"

Road Surface Condition: It is assumed that the -1 (missing) values correspond to value = 1, which according to the dictionary of definitions provided by the UK open data website corresponds to "Dry"

Junction Detail: It is assumed that the -1 (missing) values correspond to value = 0, which according to the dictionary of definitions provided by the UK open data website corresponds to "Not at junction or within 20 meters"

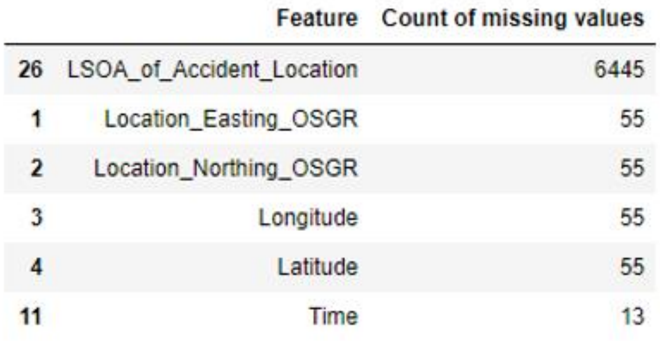
2nd Road Number: It is not of interest in this analysis, so the field is removed

Weather Conditions: It is assumed that the -1 (missing) values correspond to value = 9, which according to the dictionary of definitions provided by the UK open data website corresponds to "Unknown"

Did Police Officer Attend Scene of Accident: It is not of interest in this analysis, so the field is removed

Urban or Rural Area: It is assumed that the -1 (missing) values correspond to value = 3, which according to the dictionary of definitions provided by the UK open data website corresponds to "Unknown"

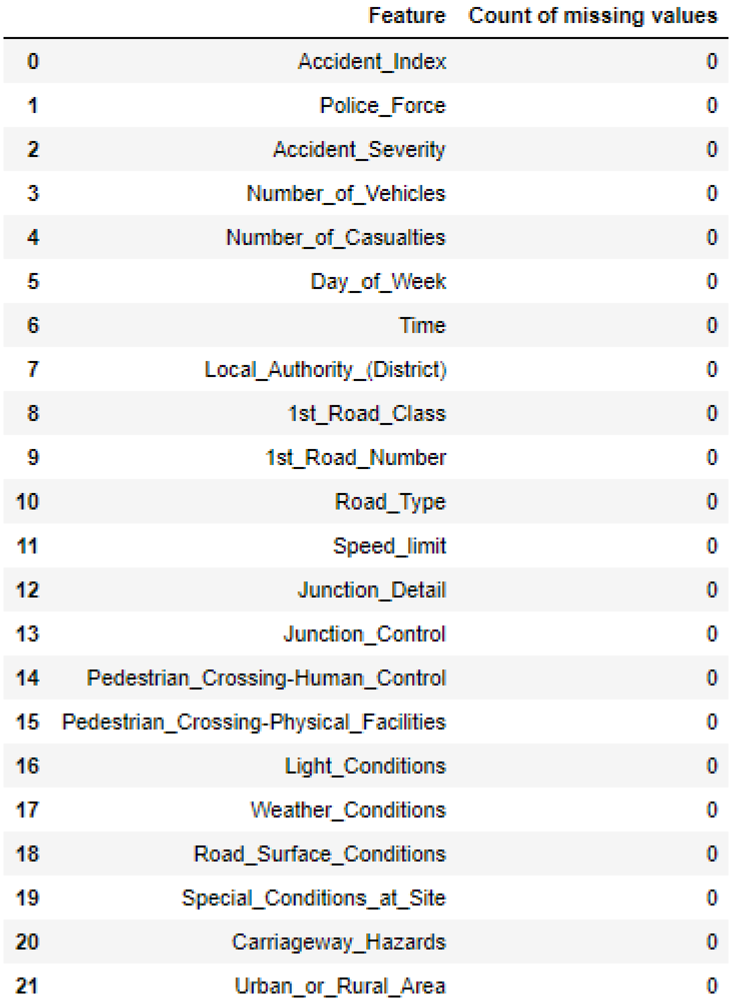
Once the effect is minimized, we proceed to review the entire dataset to know how many data is missing:



	Feature	Count of missing values
26	LSOA_of_Accident_Location	6445
1	Location_Easting_OSGR	55
2	Location_Northing_OSGR	55
3	Longitude	55
4	Latitude	55
11	Time	13

In order to deal with data missing LSOA_of_Accident_Location, Location_Easting_OSGR , Location_Northing_OSGR, Longitude and Latitude will be dropped because all of them are not of interest in this analysis, so the fields will be removed and data missing in column "Time" will be replace by 00:00.

Once the above is done, we have:



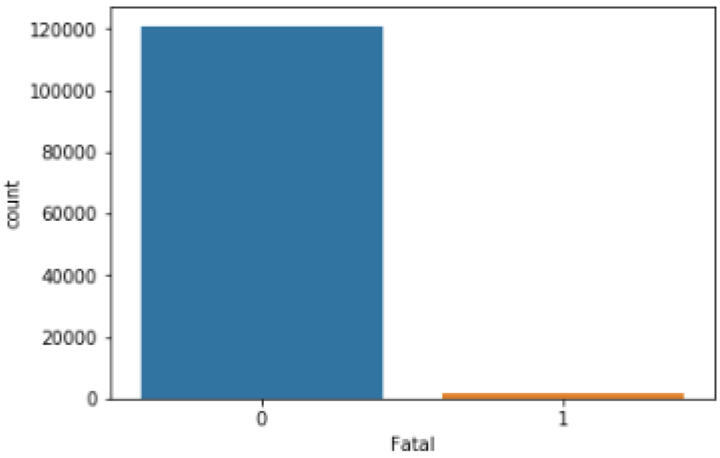
	Feature	Count of missing values
0	Accident_Index	0
1	Police_Force	0
2	Accident_Severity	0
3	Number_of_Vehicles	0
4	Number_of_Casualties	0
5	Day_of_Week	0
6	Time	0
7	Local_Authority_(District)	0
8	1st_Road_Class	0
9	1st_Road_Number	0
10	Road_Type	0
11	Speed_limit	0
12	Junction_Detail	0
13	Junction_Control	0
14	Pedestrian_Crossing-Human_Control	0
15	Pedestrian_Crossing-Physical_Facilities	0
16	Light_Conditions	0
17	Weather_Conditions	0
18	Road_Surface_Conditions	0
19	Special_Conditions_at_Site	0
20	Carriageway_Hazards	0
21	Urban_or_Rural_Area	0

Having fixed missing values and values encoded as -1, I am going to balance the dataset.

2.3. Data Balance.

To make the selection of variables it is necessary to normalize some information and balance the dataset. To balance the dataset, first its necessary to normalize some features, in this case, the severity levels which is our main interest, this will be done by isolating each level of severity with typical normalizations, where the values of the new variable oscillate between 0 and 1. Let´s take only Fatal Accidents to check the balance or imbalance . Once normalization is done, we have to class:

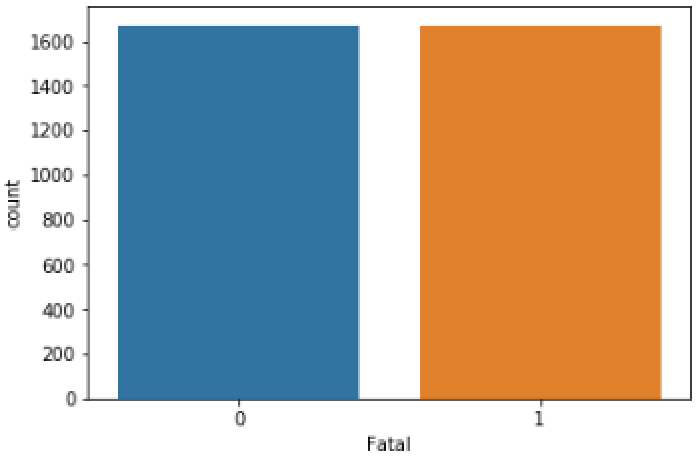
- 1 = means Fatal accident (Yes)
- 0 = means No Fatal accident (No)



As we can see, the data set is unbalanced, 1.36% of the total corresponds to fatal accidents and 98.63% are not. In this way it is necessary to balance the information, so the models to be built will not biased by this proportion.

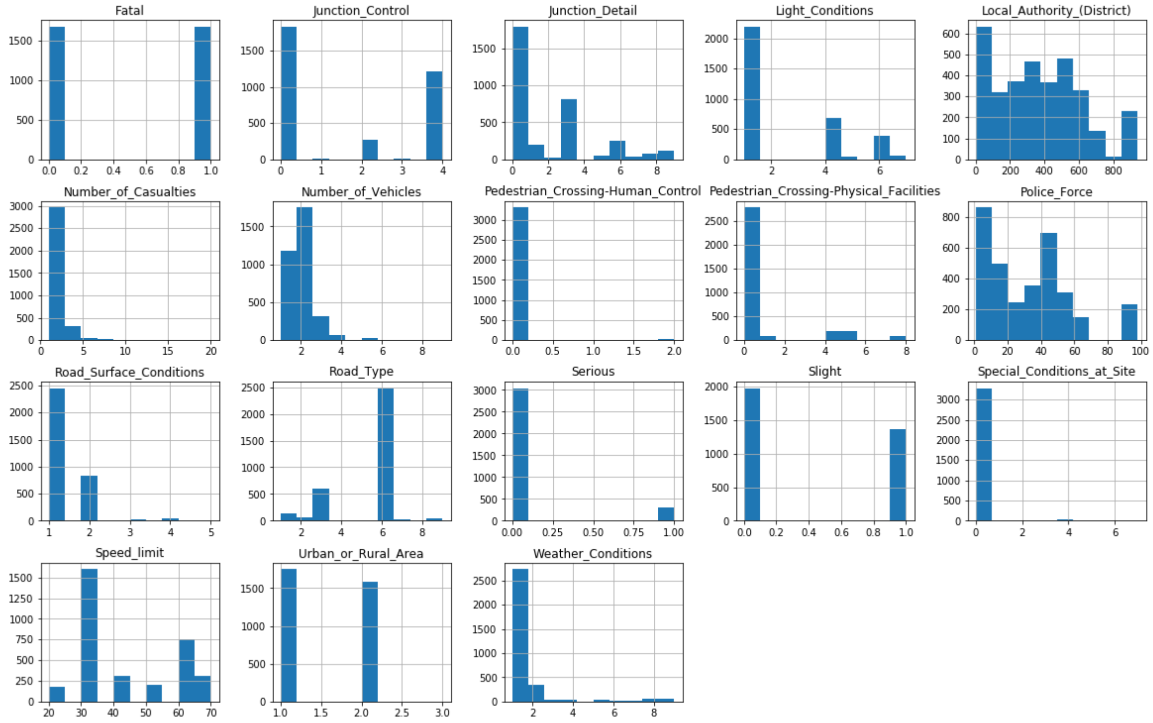
There are lots of options at this stage. Its possible to resampling the dataset by oversampling the fatalities, or under sampling the severe and slight accidents. We could train simpler models using an area under the ROC curve metric, or we could use tree-based models like gradient boost or random forest

I decide to go to undersampling. The first step is to find the quantity of the minority class, in this case 1671 samples, then find the index of the majority class when the class is 0, in such a way that the random sampling takes only these indices and randomly choose the same amount of minority class sample. Then concatenates the minority class with the samples of the majority class and finally graph the new frequency of the new dataset



2.4. Feature Selection.

Once the dataset is balanced let's visually inspect to see what we can learn from it before we begin building models.



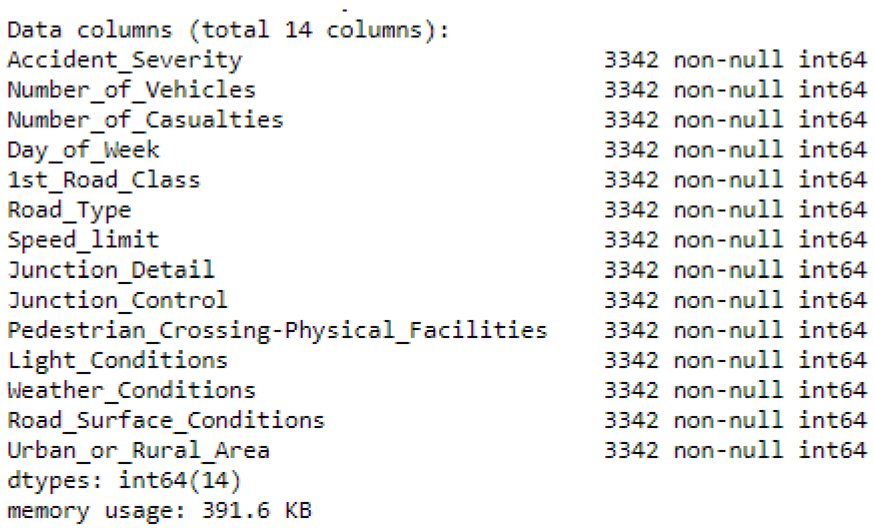
Now, we could see from some of these histograms that some of the data is still quite skewed. It is the case of:

Carriageway_Hazards is almost always zero, suggesting that fatal accidents do not occur frequently from road hazards like Vehicle load on road, Other object on road, Previous accident or even Dog or other animal on road, among others. It is therefore unlikely to be predictive.

Pedestrian_Crossing-Human_Control is almost always zero, suggesting that fatal accidents do not occur frequently at crosswalks, therefore it is unlikely to be predictive.

Special_Conditions_at_Site is almost always zero, which suggests that fatal accidents do not occur frequently due to Auto traffic signal - out, Auto signal part defective, Road sign or marking defective or obscured, Roadworks, Road surface defective, Oil or Diesel. It is therefore unlikely to be predictive.

The next step is select a series of factors that are likely to be predictive, eliminating some of the descriptive variables such as Police_Force, Local_Authority (District), 1st_Road_Number and the fields of each level of severity among other, so the selected features will be:



```
Data columns (total 14 columns):
Accident_Severity                           3342 non-null int64
Number_of_Vehicles                          3342 non-null int64
Number_of_Casualties                        3342 non-null int64
Day_of_Week                                 3342 non-null int64
1st_Road_Class                              3342 non-null int64
Road_Type                                   3342 non-null int64
Speed_limit                                 3342 non-null int64
Junction_Detail                             3342 non-null int64
Junction_Control                            3342 non-null int64
Pedestrian_Crossing-Physical_Facilities     3342 non-null int64
Light_Conditions                            3342 non-null int64
Weather_Conditions                          3342 non-null int64
Road_Surface_Conditions                     3342 non-null int64
Urban_or_Rural_Area                         3342 non-null int64
dtypes: int64(14)
memory usage: 391.6 KB
```

For the analysis and creation of a predictive model and because the dataset is labeled, I am going to use supervised machine learning. Now the information will be analyzed to determine which of the attributes have the highest correlation with the target variable **Accident Severity.**

The variables of the dataset will be group into three main sets:

Accident details	**Location and time**	**Environmental issues**
1. Accident Severity 2. Number of Vehicles 3. Number of Casualties 4. 1st Road Class 5. Road Type 6. Speed limit 7. Junction Detail 8. Junction Control	1. Day of Week 2. Urban or Rural Area	1. Light Conditions 2. Weather Conditions 3. Road Surface Conditions 4. Pedestrian Crossing Physical Facilities

With all the data above, it is possible to determine interesting situations and make questions about them, among others:

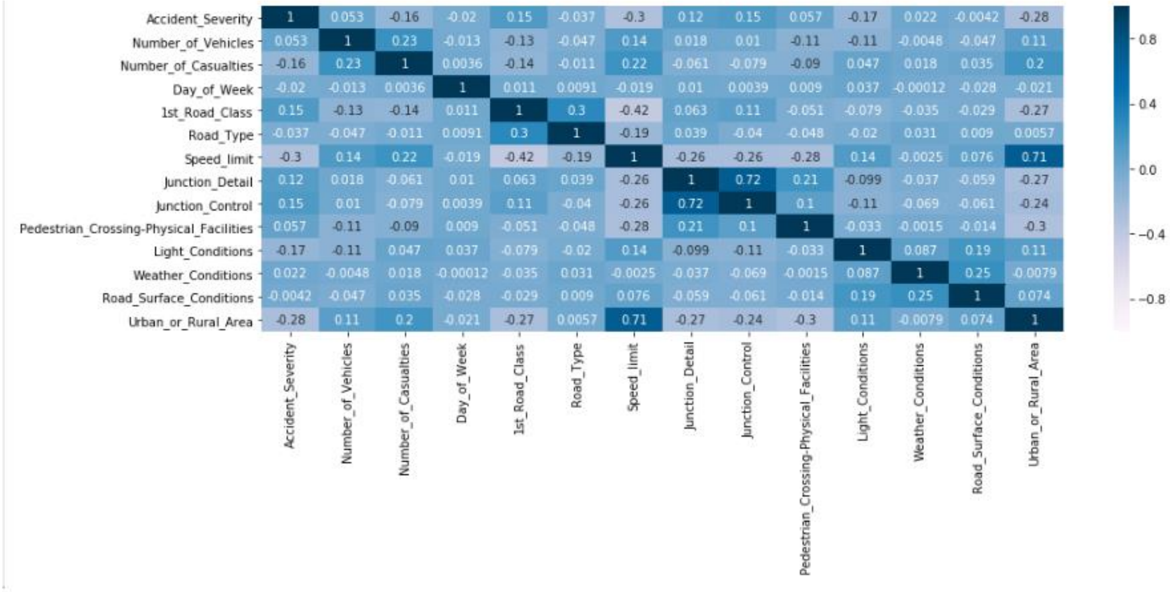
1. When do most accidents happen? On weekdays or the weekend?
2. Is weather a key factor in road accidents?
3. Which causes more accidents?

3. EXPLORATORY DATA ANALYSIS (EDA)

3.1. Methodology

Since we are interested in a causal relationship, it makes sense to go even deeper individually into the other variables and remove the last ones from the dataset.

After cleaning the data, we finally have 3.342 rows and 14 characteristics correlated with Fatal Accident for modeling, from which it is possible to extract important information to understand the behavior of the target variable.



Now we would like to know the significant of the correlation between accident severity and the other variables

3.2. Relationship between Accident Severity and Number of Vehicles Involved

Regarding the severity of the accident due to the number of vehicles involved, the hypothesis here is that the greater the number of cars involved, the greater the severity of the accident should be, however, according to the distribution of the information, the number of fatal accidents is almost the same. when the accident involves only one car than when it involves two. The same behavior occurs with serious accidents, not so with minor accidents in which the number with two cars involved doubles the number of accidents with a single vehicle. Since the p-value is < 0.05, we say there is moderate evidence that the correlation is significant between Number of Vehicles and Accident Severity although the linear relationship is isn't extremely strong (~0.050) the two variables most likely do not affect each other.

-

3.3. Relationship between Accident Severity and Number of Casualties

The hypothesis here is that the greater the number of victims, the greater the severity of the accident, however, according to the distribution of the information, this hypothesis is not correct and is verified in the statistical results. Since the p-value is <0.001, we say there is strong evidence that the correlation is significant between Number of Casualties and Accident Severity although the linear relationship is negative and isn't extremely strong (~ 0.16).

The number of fatal accidents is almost equal to the number of light accidents and triples the number of severe accidents when there is a single person involved, when there are two or more the number of fatal accidents decreases to almost a quarter of the number.

3.4. Relationship between Accident Severity and Day of Week

Since the p-value is >> 0.1, we say there is no evidence that the correlation is significant between Day of Week and Accident Severity, the linear relationship is negative and isn't extremely strong (~0.02) so the two variables most likely do not affect each other.

3.5. Relationship between Accident Severity and 1st Road Class

The hypothesis here is that the better category of road, accidents should be more serious, however, according to the distribution of the information, this hypothesis is not correct, the highest number of fatal and light accidents occurs on the roads type A (roads with two lanes in each direction separated by small level), while fatal and light accidents are very few on motorways. Since the p-value is << 0.001, we say there is strong evidence that the correlation is significant between 1st Road Class and Accident Severity although the linear relationship isn't extremely strong (~0.14)

3.6. Relationship between Accident Severity and Road Type

The hypothesis here is that the type of road (roundabout, one-way street, highway etc ...) should have an impact on the number of accidents, probably roundabouts will be the type of road where accidents occur more frequently due to high traffic flow However, the hypothesis is false, according to the data set where more fatal and light accidents occur is in Single carriageway. Statistically the p-value is << 0.05, we say there is moderate evidence that the correlation is significant between Road Type and Accident Severity although the linear relationship is isn't extremely strong (~ 0.040) and the two variables most likely do not affect each other.

3.7. Relationship between Accident Severity and Speed limit

The hypothesis here is that when exceeding the speed limit there is a positive linear relationship, where the higher the speed the greater the severity of the accident, however the hypothesis is not correct, according to the data set exceeding the speed limit of 30 mph and 60 mph, generates many more fatal and minor accidents than exceeding the speed limit of 70 mph or more. Statistically the p-value is <0.001, we say there is strong evidence that the correlation is significant between Speed Limit and Accident Severity although the linear relationship isn't strong (~ 0.3)

3.8. Relationship between Accident Severity Junction Detail and Junction Control

The hypothesis here is that the detail and the control of the junction influences the accident rate. Statistically the p-value is <0.001, for Junction Detail and for Junction control, so we can say there is strong evidence that the correlation is significant between Junction Detail, Junction Control and Accident Severity although the linear relationship isn't strong in both cases.

3.9. Relationship between Accident Severity and Pedestrian Crossing Physical Facilities

The hypothesis here is that the type of pedestrian crossing facility could influence the accident rate. It could be thought that where there is no pedestrian crossing control, the accident rate is higher and effectively with the analysis of the data set it is found that the greatest number of fatal, serious or light accidents occur where there is no pedestrian crossing control.

Since the p-value is $<$ 0.05, we say there is moderate evidence that the correlation is significant between Pedestrian Crossing Physical Facilities and Accident Severity although the linear relationship isn't strong (~0.056 and the two variables most likely do not affect each other.

3.10. Relationship between Accident Severity and Light Conditions

The hypothesis here is that the lighting conditions have a high impact on the accident rate, specifically it would be expected that as the light decreases, the number of accidents increases, however this hypothesis is wrong, according to the distribution of the data, the greater amount of Fatal, serious and light accidents occur where there is still daylight followed by conditions where there is Darkness - lights lit.

Since the p-value is < 0.001, we say there is strong evidence that the correlation is significant between Junction Detail and Accident Severity although the linear relationship isn't strong (~0.1)

3.11. Relationship between Accident Severity and Weather Conditions

The hypothesis here is that the weather conditions have an impact on the accident rate, specifically it would be expected that in conditions such as rain, snow or strong winds, the number of accidents will increase, however this hypothesis is wrong, according to the distribution of the data the greater amount of fatal, serious and light accidents occur with optimal weather conditions and no wind.

Since the p-value is > 0.1, we say there is no evidence that the correlation is significant between Weather Conditions and Accident Severity. The linear relationship isn't extremely strong (~0.02) and the two variables most likely do not affect each other.

3.12. Relationship between Accident Severity and Road Surface Conditions

The hypothesis here is that the conditions of the road surface directly influence the accident rate, this means that on Wet or damp, with snow, Frost or ice, Flood over, Oil or diesel or even mud roads the accident rate is Higher than on roads with dry surfaces, however statistically this statement is not correct. According to the distribution of the data, the greater number of accidents, regardless of the level of severity, occur in dry road conditions. Since the p-value is> 0.1, we say there is no evidence that the correlation is significant between Road Surface Conditions and Accident Severity.

3.13. Relationship between Accident Severity and Urban or Rural Area

The hypothesis here is that the number of fatal accidents should be higher in urban areas than in rural areas, due to the greater number of cars, pedestrians or other variables, however this hypothesis is not correct, the greater number of fatal accidents occurs in rural areas and the greatest number of light accidents occurs in urban areas, it could be thought that excess speed is greater in rural areas, which would explain the greater fatality. Statistically, the p-value is <0.001, so we can say there is strong evidence that the correlation is significant between Urban or Rural Area and Accident Severity although the linear relationship isn't strong (~ 0.3)

Now here we could choose between the three levels of accident severity in order to develop the model, so I rather to choose the one with the most impact on health, in this case fatal accidents. We now have a better idea of what our data looks like and which variables are important to consider when predicting the accident severity. We have narrowed it down to the following variables:

- **Number of Casualties**
- **1st Road Class**
- **Speed limit**
- **Junction Detail**
- **Junction Control**
- **Light Conditions**
- **Urban or Rural Area**

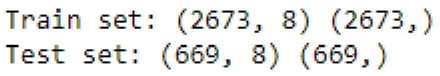
With this feature we are going to build the model.

4. MODELING

Most accidents happen on A roads or unclassified roads. Most accidents happen at low speed limits. These findings, and similar ones across the variables, likely reflect the most commonly occurring scenarios (i.e. drivers travelling at low speeds, on the most frequently occurring road types, during the day, in good weather and on good road surfaces). It seems that accidents happen most often in conditions that occur most often, which perhaps is not surprising. Consequently, we expect it will be challenging to build a predictive machine learning model, as it is likely there is limited signal in this data, and lots of noise.

We then one-hot encoded the categorical variables and then built a series of different machine learning models to predict the accident severity, with varying degrees of success. The accuracy scores of each approach are shown below.

After selecting the feature set X and y and normalize X, its time to use the test-train split and select a model to use. The distribution set is: 80 % train, 20% set, according to this we have:



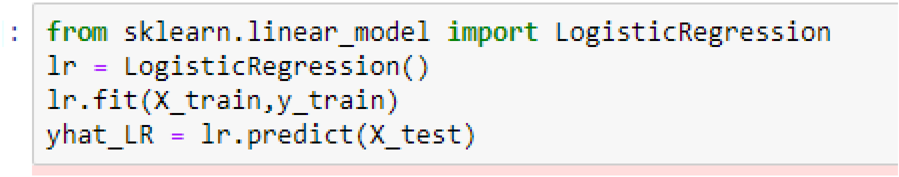
```
Train set: (2673, 8) (2673,)
Test set: (669, 8) (669,)
```

For our purpose, classification models are the best option they can provide the probabilities of an accident may be fatal or not. In order to find the best model, let´s select some types:

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machine (SVM)

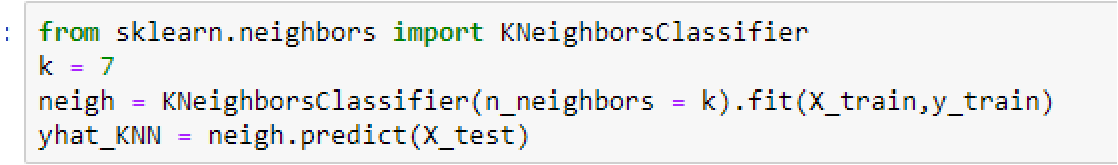
With the following structures:

4.1. Logistic Regression



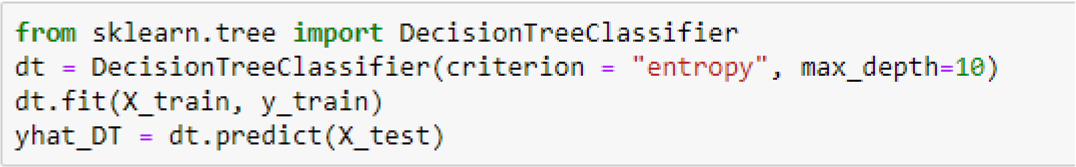
```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train,y_train)
yhat_LR = lr.predict(X_test)
```

4.2. K-Nearest Neighbors



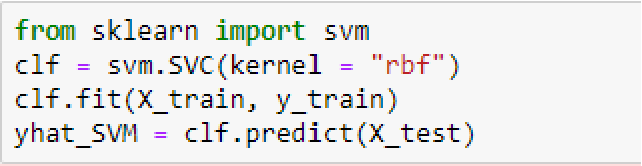
```
from sklearn.neighbors import KNeighborsClassifier
k = 7
neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
yhat_KNN = neigh.predict(X_test)
```

4.3. Decision Tree



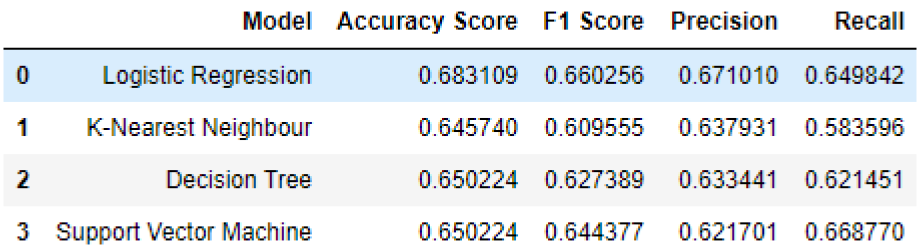
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion = "entropy", max_depth=10)
dt.fit(X_train, y_train)
yhat_DT = dt.predict(X_test)
```

4.4. Support Vector Machine (SVM)



```
from sklearn import svm
clf = svm.SVC(kernel = "rbf")
clf.fit(X_train, y_train)
yhat_SVM = clf.predict(X_test)
```

4.5 Performance and Results



	Model	Accuracy Score	F1 Score	Precision	Recall
0	Logistic Regression	0.683109	0.660256	0.671010	0.649842
1	K-Nearest Neighbour	0.645740	0.609555	0.637931	0.583596
2	Decision Tree	0.650224	0.627389	0.633441	0.621451
3	Support Vector Machine	0.650224	0.644377	0.621701	0.668770

Note that accuracy is a relatively the same for all models however the best one is the logistic regression, as the class is balance means that a model can score well solely by predicting which accidents belong to the largest class.

5. CONCLUSION

Although we have had some success in building a model, it is unlikely that any of these would be usable in practice, and further work needs to be done on feature engineering and setting up the problem to build a more robust model. It is likely though that better results could be achieved by recognising that accidents tend to happen where people/vehicles are most frequently present, so setting up the problem instead as an anomaly detection problem rather than a classification problem, to look at deviations from this pattern, may lead to better results.

Undoubtedly, a model with this characteristic must be working with real time data otherwise is just useful information, but with the proper data this approach could safe a lot of life or avoid a lot of accidents.