Capstone Project - Car accident severity

**Description & Discussion of the Background**

Road Accident is the most undesirable and unexpected thing to occur to a road user, though they happen quite often. Unfortunately, we can see a minatory rise of road accidents in United Kingdom, conspicuously highroad accidents over the past few years. It has a massive impact on society as well as in the economy of our country as there is an immense cost of fatalities and injuries. According to a recent report, annually on an average 9,000 lives have been taken by road accidents and lead to almost 28,000 injuries. This record indicates that every day, approximately 28 people were killed by road accidents and it is quite devastating. Besides this, according to WHO, the economic cost of road accidents to a developing country like us is 2-3% of GDP, which is a significant loss for a country like ours. Moreover, reducing this loss has become a great matter of concern for our country now.

**Data Description**

For the accurate prediction of the severity of accidents, a considerable number of traffic accident records with full information is required to train by using the proposed approaches. In this research work, the authors have collected a dataset from the Traffic Bureau that consists of total 37,885 traffic accidents record from the year 2007-2017. The entire dataset will split into two parts- Training Dataset and Test Dataset. 70% of the whole dataset has been chosen randomly by using a python library as a training data set and the remaining 30% has been used as our test dataset. We have used the 70-30 ratio for splitting dataset because of its proven accuracy.

**Audience**

People who travel on a daily basis or even if someone new commuting to the highway will have an idea of the road condition and how time, weather and other scenarios while driving may cause collision. They will get awareness to take precautions

**Benefits**

it will reduce the number of collision since they will have better idea of what to expect

**Data**

The data we will be using is the collision data from the example, as well as the pdf describing the data, for reference

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

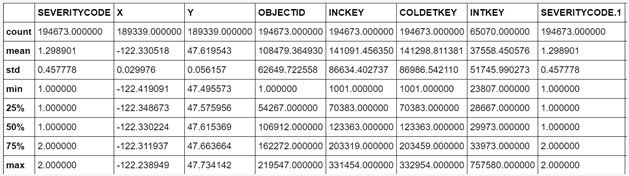
**Methodology**

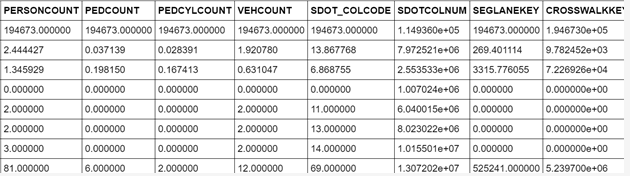
#### Initial data will be cleaned in Excel and then using Python for additional cleaning, analyzing and visualizing state traffic accident data. The idea is to extract data to evaluate its finding, combine factors leading to collision together, and building linear regression models to evaluate the chances of collision.

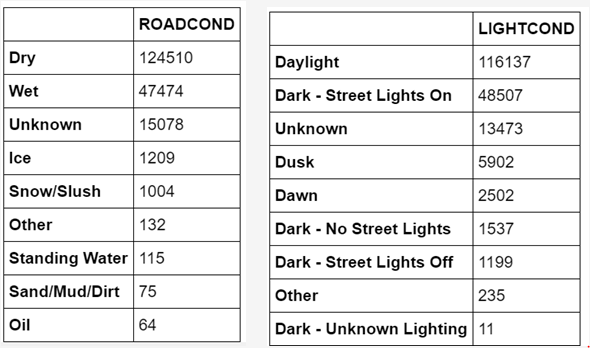
To do this, we will be using the following modules/libraries

* Pandas
* Numpy
* Matplotlib
* Scikit-learn

As a database, I used GitHub repository in my study. My master data which has the main components Severity Code, Weather Elements, Light Condition, Junction Type and Collission Type. For classification problems, matplotlib is extensively used the supervised algorithm. The primary perspective of this algorithm is predicting the value of the desired variable by learning decision rules deduced from the features of the data and create a model of that. A root node is designated for the construction of this model based on the best attribute picked by the gain approach and the sub-nodes are then generated on the basis of the decision taken in relation to the status of quality selected at each node. When each node is reduced to a single quality status, the class is determined at the end of the node; it is called a leaf. These courses of action continue recursively until a class is defined at the end of each node.







**Results**

In this research paper, I have determined the performance of each algorithm, for four accident severity classes (Fatal / Grievous /Simple Injury/ Motor Collision). By overall performance, Gradient-Boost and Random Forest gives the best result because of its iterative classification on matplotlib.



**Discussion**

We observe that most of the accidents in our dataset are Fatal and value for the other three classes is very low. For that reason, in second experiment, we merge Grievous, Simple Injury, Motor Collision these three accident severity classes into one class. Therefore, we have attained the performances of the proposed approaches for two accident severity classes (Fatal / Grievous). In this experiment, we have noticed that the accuracy of interpolated data get increased and remain the same. But it is also mentionable that, the performance is much better than the previous experiment as precision and F1 score increased here in a noticeable way. Besides this, we did experiment with the features in our dataset and have tried to find out their effect on a traffic accident. Statistically I have found that based on the condition of some features the number of accidents gets increased. It’s a significant noticeable thing for making proper steps to decrease the number of accidents.

**Conclusion**

As a result, people are turning to big cities to start a business or work. From the above table, we can see that servere car accidents occurs frequently under clear and dry condition at intersection. Besides, speeding is also an important factor leading the accident happen.