PREDICTING TRAFFIC ACCIDENT SEVERITY

Introduction:

Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally and are predicted to become the seventh leading cause of death by 2030. Analysing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed. These insights, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accident can occur as well as saving both, time and money.

In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance. Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

Objective:

The objective of the project is to use a dataset with conditions and the severity of the reported occurrence of car accidents in a city to predict the severity label which describes the fatality of an accident given the conditions. The conditions such as weather, light, speeding, inattention and user information is used for training and forecasting the severity and occurrence of accident.

Data:

The dataset is the collisions data in Seattle provided by SPD and recorded by Traffic records. This includes all types of collisions and is updated weekly. The dataset consists of 37 attributes which describe the location geometry, road conditions and user information.

The attributes describe the location of the collision, collision type, total number of people involved, number of pedestrians involved, number of vehicles involved, total injuries involved, number of serious injuries in the collision, number of fatalities, date and time of the incident, category of the junction, code of the collision and description of the code, whether the collision was due to inattention, whether the driver was under the influence of drugs, the weather conditions during the time of collision, the road conditions during the time of collision, the light conditions during the time of collision, whether the pedestrian was in the right of way, whether speeding was a factor, the state provided code for the collision and it’s description, the lane segment of the collision, the crosswalk at the collision and whether the collision involved in hitting a parked car.

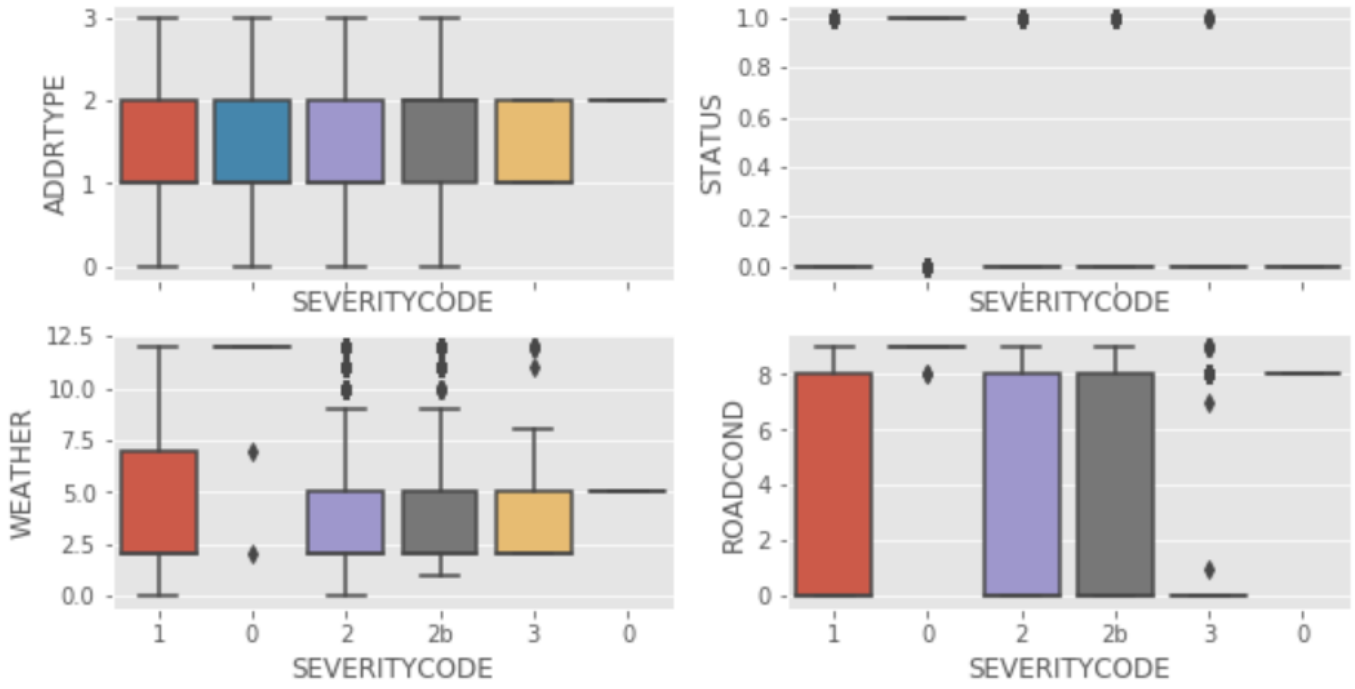
The label is the severity of the collision and it describes the level of severity from 0 – 3. Where 0 is unknown, 1 is for property damage, 2 is for injury, 2b is for serious injury and 3 is for fatality. This data can be used in taking precautionary measures.

For example, if the supervised learning predicts that the accidents at the specific crosswalk or location is highly fatal, warning signs or alternative pedestrian walkways can be installed.

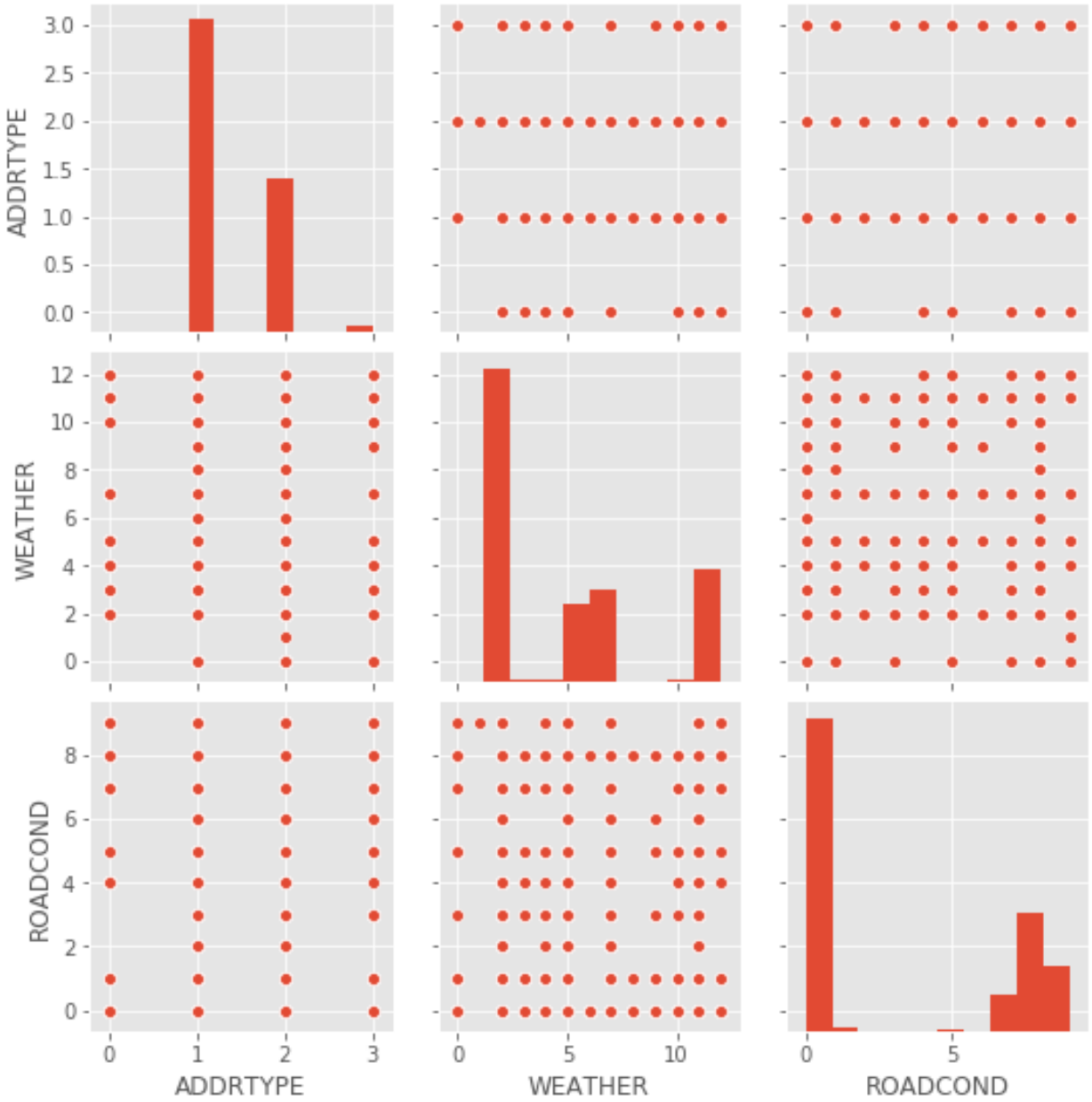
Exploratory Data Analysis:

Exploratory analysis is done to identify the correlation among the parameters and the parameter and the label.

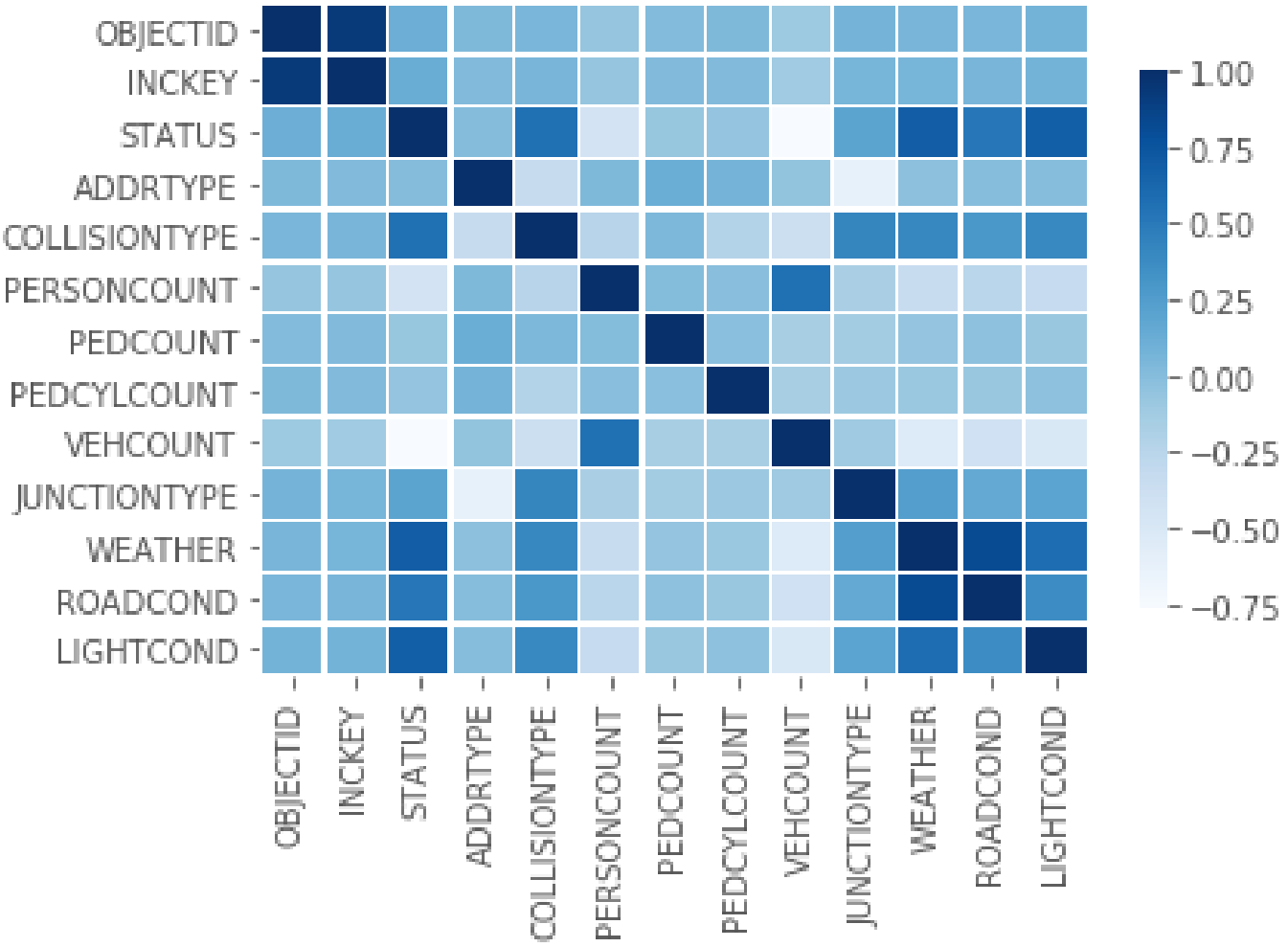
The data set is first standardized and then plotted using a box and whisker plot as shown in figure which is a plot used to graphically display patterns of quantitative data. It is convenient to use a box plot with multiple variables as it can graphically show the variation within the ranges and the variation between the variables. Plotting a box plot for the severity with respect to the ADDRTYPE, STATUS, WEATHER, ROADCOND shows that the severity of the accident varies more across the weather, but it is uniform otherwise.



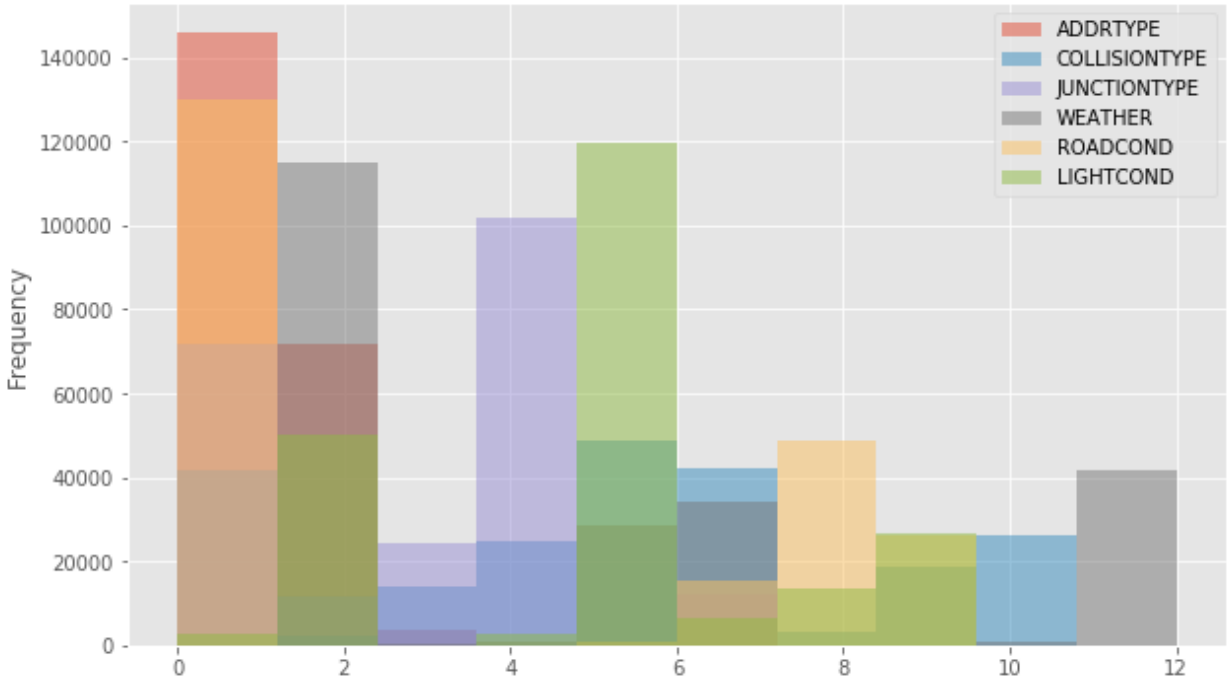
The pair plot is plotted to determine the covariance between the address type, weather and the road condition. All the scatter plots in the pair plots does not show a positive linear tendency and therefore it confirms that the data is not normally distributed.



The covariance matrix heatmap shows whether there is a positive or negative correlation between the attributes. The weather and road condition have a high positive correlation.



A histogram for the address type, collision type, junction type, weather, road condition, light condition attributes are plotted below.



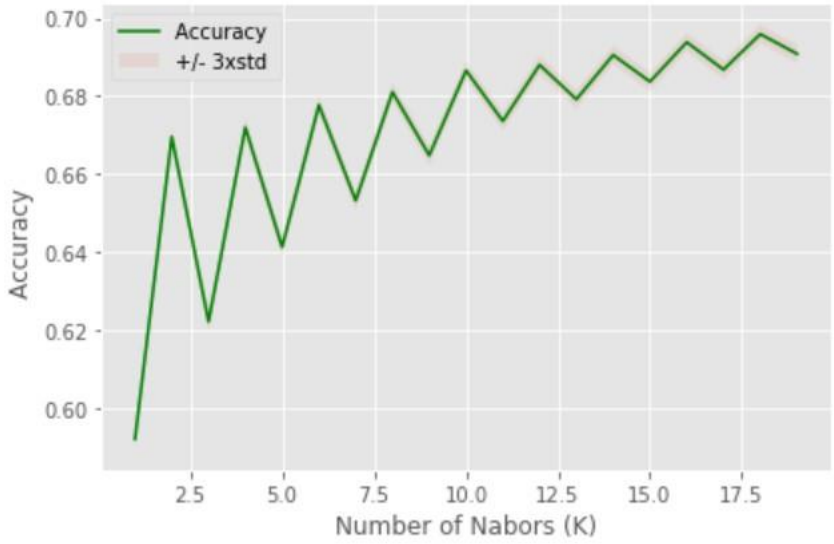
Predictive Modelling:

The model was built with 80% train size and 20% test size with a random state of 1. The algorithms used were K-Nearest Neighbors, Decision tree, Logistic regression, Naive Bayes and Random forest and the accuracy score were tested using the confusion matrix.

K-Nearest Neighbors:

K-nearest neighbor learning is an instance-based learning method which learns by simply storing the training examples rather than constructing an explicit description of the target function. This type of classification is used in storing and reusing past experiences. The standard Euclidean distance is used to define the nearest neighbors of an instance.

From the below image, the maximum accuracy can be obtained at K=17.



The accuracy of the algorithm on our dataset from the confusion matrix is as below,

```
Score: 0.6867856684217285
F1-Score: 0.8104688932751037

Confusion Martix:
[[26074  1285]
 [10910   666]]

Classification Report:
              precision    recall  f1-score   support

           1       0.71      0.95      0.81     27359
           2       0.34      0.06      0.10     11576

    accuracy                           0.69     38935
   macro avg       0.52      0.51      0.45     38935
weighted avg       0.60      0.69      0.60     38935
```

Decision Tree:

The basic algorithm used in this is the ID3 algorithm which builds the tree using a top-down, greedy approach. This works by selecting an attribute and assigning a test case for it. For each test case results, a new descendant node is created. This is iterated until the training examples are sorted to the appropriate descendant leaf nodes and are perfectly classified.

The accuracy of the algorithm on our dataset from the confusion matrix is as below,

```
Score: 0.7506613586747143
F1-Score: 0.84807036213966

Confusion Martix:
[[27095   264]
 [ 9444  2132]]

Classification Report:
              precision    recall  f1-score   support

           1       0.74      0.99      0.85     27359
           2       0.89      0.18      0.31     11576

    accuracy                           0.75     38935
   macro avg       0.82      0.59      0.58     38935
weighted avg       0.79      0.75      0.69     38935
```

Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The accuracy of the algorithm on our dataset from the confusion matrix is as below,

```
Score: 0.6984718119943496
F1-Score: 0.8216997752262925

Confusion Martix:
[[27052   307]
 [11433   143]]

Classification Report:
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     27359
           2       0.32      0.01      0.02     11576

    accuracy                           0.70     38935
   macro avg       0.51      0.50      0.42     38935
weighted avg       0.59      0.70      0.58     38935
```

Naïve Bayes:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

The accuracy of the algorithm on our dataset from the confusion matrix is as below,

```
Score: 0.701476820341595
F1-Score: 0.8233720841881317

Confusion Martix:
[[26074  1285]
 [10910   666]]

Classification Report:
              precision    recall  f1-score   support

           1       0.70      0.99      0.82     27359
           2       0.45      0.02      0.04     11576

    accuracy                           0.70     38935
   macro avg       0.58      0.50      0.43     38935
weighted avg       0.63      0.70      0.59     38935
```

Random Forest:

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks).

The accuracy of the algorithm on our dataset from the confusion matrix is as below,

```
Score: 0.7164504944137665
F1-Score: 0.8060163058757379

Confusion Martix:
[[26074  1285]
 [10910   666]]

Classification Report:
              precision    recall  f1-score   support

           1       0.78      0.84      0.81     27359
           2       0.53      0.43      0.47     11576

    accuracy                           0.72     38935
   macro avg       0.65      0.63      0.64     38935
weighted avg       0.70      0.72      0.71     38935
```

Results:

From the below table, it can be said that the best classifier of this problem is Decision Tree. It's gotten best score and more true positive values.

Classifier	Score	F1 Score
KNN	0.68	0.81
Decision Tree	0.75	0.84
Logistic Regression	0.69	0.82
Naïve Bayes	0.70	0.82
Random Forrest	0.71	0.80

Conclusion:

A traffic collision, also called a motor vehicle collision, car accident, or car crash, occurs when a vehicle collides with another vehicle, pedestrian, animal, road debris, or other stationary obstruction, such as a tree, pole or building. Traffic collisions often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. Purpose of this project was to identify the type of collisions or type of injury to the during of collision. The major important of predicting is Weather condition, Road condition, Address of collision, how many peoples are involved, how many vehicles are present and Which type of vehicles. That are helps to predicted to what type of injury or disability in collision.