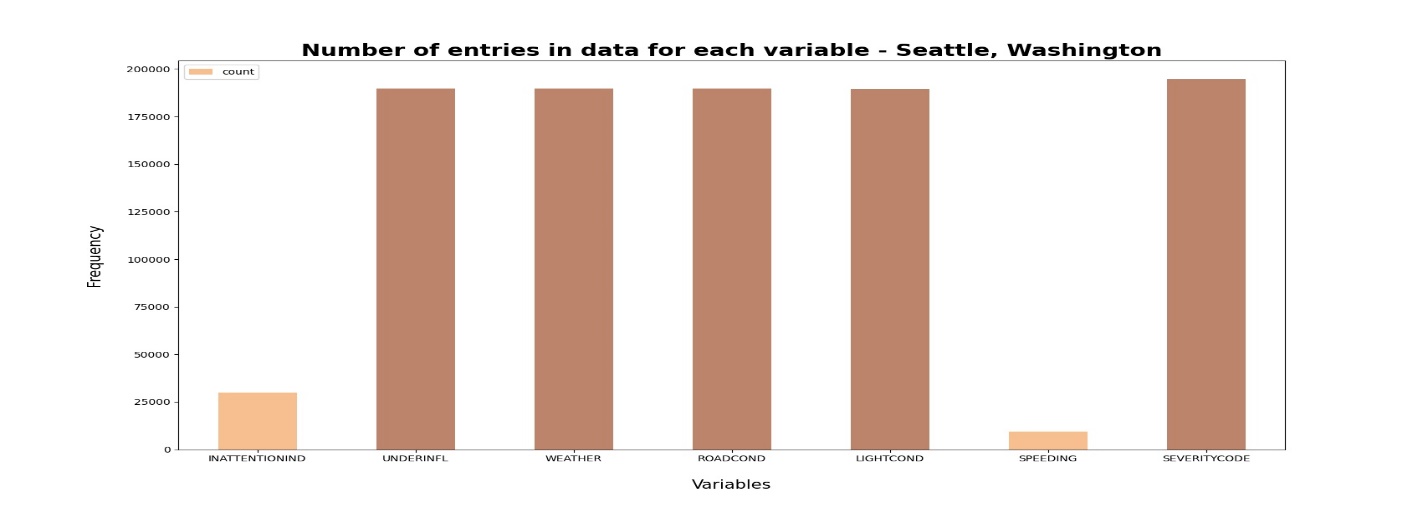
**Introduction/Business Problem**

The science and technology has changed the world transportation sector. Now people love to travel in cars intercity or long distances with greater comfort and speed. However, the problem of car accidents are still suffer us a lot. The USA Highway and Traffic Administration department said there is a loss of about 900 billion dollars in a year due to car accidents besides loss precious human lives. In 2017, Washington traffic fatalities increased by 5.4 percent (from 536 to 565). Nationally, traffic deaths remained largely unchanged, decreasing less than half a percent and leveling off the steepest two-year increase in 50 years. With the rebounding economy and increases in vehicle miles traveled, Washington must require to implement new innovations and modern data analyzation to realize its vision of zero traffic fatalities and serious injuries. This Data Science Project aims to analyze the traffic accidents data and find the factors that caused these accidents. The main stake holders involved in this are 1- Public Development Authority Seattle 2- Drivers

**Data:**

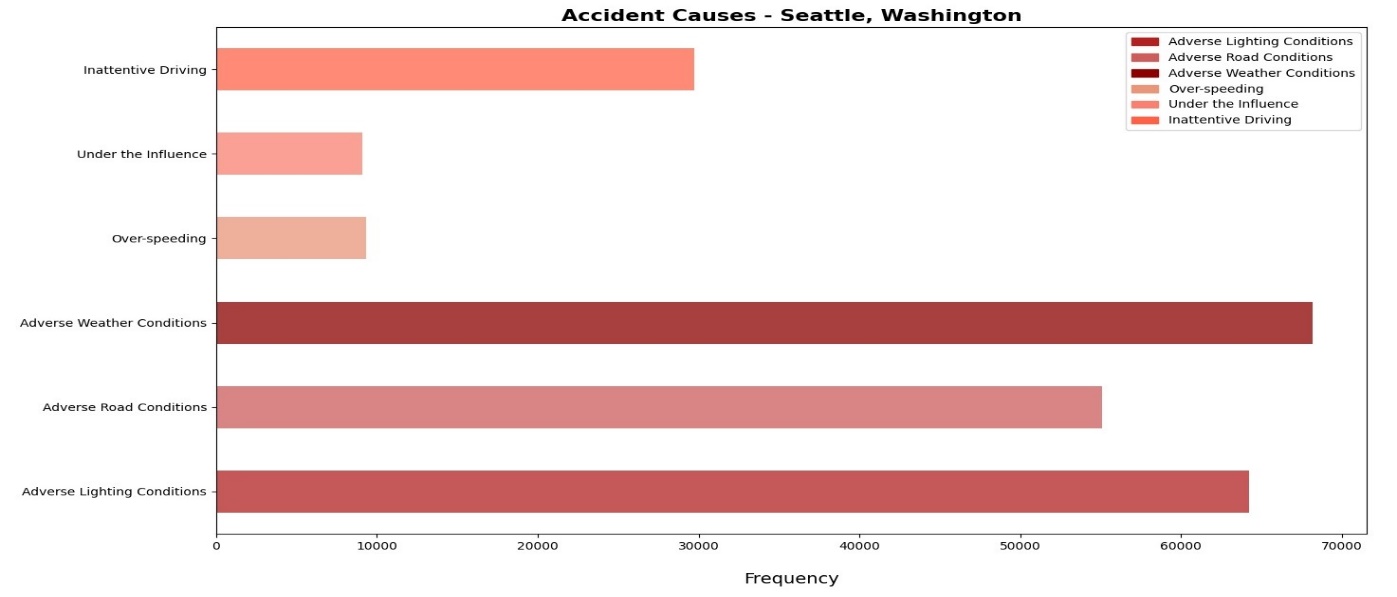
The latest accident dataset of accidents has been taken from the city of Seattle, Washington from 2004 to 2020. The data contains accidents details such as time, condition of weather, location, severity, junction type, road conditions, speeding etc. There are more than 30 features in the dataset. The data set used can be downloaded from [there.](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) The propose model will try to predict an accident severity, the feature accident severity is in the form of 1( Car or property damage) and 2 (physical injury) that are encoded into 0 (property damage only) and 1( Physical Injury) using One hot Encoder. The 1 denote higher severity as due to Physical damage than 0 which indicates property damage only. However, there would be many cases in which the condition will be very serious such as loss of lives which are unique values for every variable which were either Other or Unknown. To solve this, the arrays were made for each column which are encoded equal proportions of elements as original columns. These arrays were used on the original columns of which had Others or Unknown in them to fill the missing data. This entire process of cleaning and filling of data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier. The INATTENTIONIND (whether or not driver was inattentive (Y/N), UNDERINFL (Whether or not the driver was under the influence (Y/N), WEATHER (Weather conditions Overcast\Rain\Clear), ROADCOND (road conditions Wet or dry), LIGHTCOND(Light conditions), SPEEDING (Whether the car was above the speed limit Yes o No) are taken as the features of the Dataset.



**Methodology:**

**Exploratory Analysis**

Considering the categorical data features such as INATTENTIONIND (whether or not driver was inattentive (Y/N), UNDERINFL (Whether or not the driver was under the influence (Y/N), WEATHER (Weather conditions Overcast\Rain\Clear), ROADCOND (road conditions Wet or dry), LIGHTCOND(Light conditions), SPEEDING (Whether the car was above the speed limit Yes o No). These categorical data features have limited values and usually based on a particular finite group whose correlation might depict a different image then what it actually is. This can make our dataset unbalance. Hence, SMOTE function was used from imblearn library in order to balance the feature variable in equal proportions in order to have an unbiased classification model which is trained on equal instances of both the elements under severity of accidents.

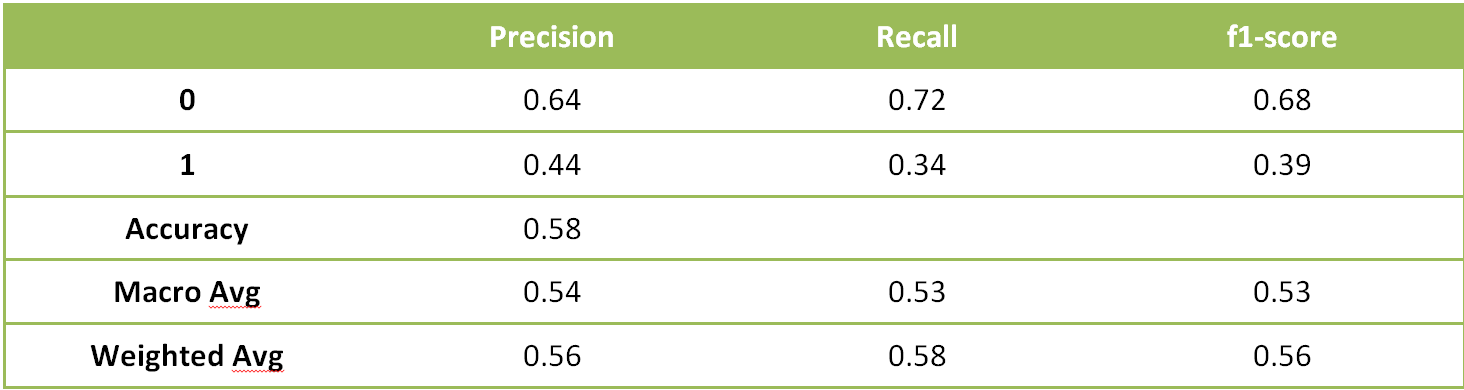


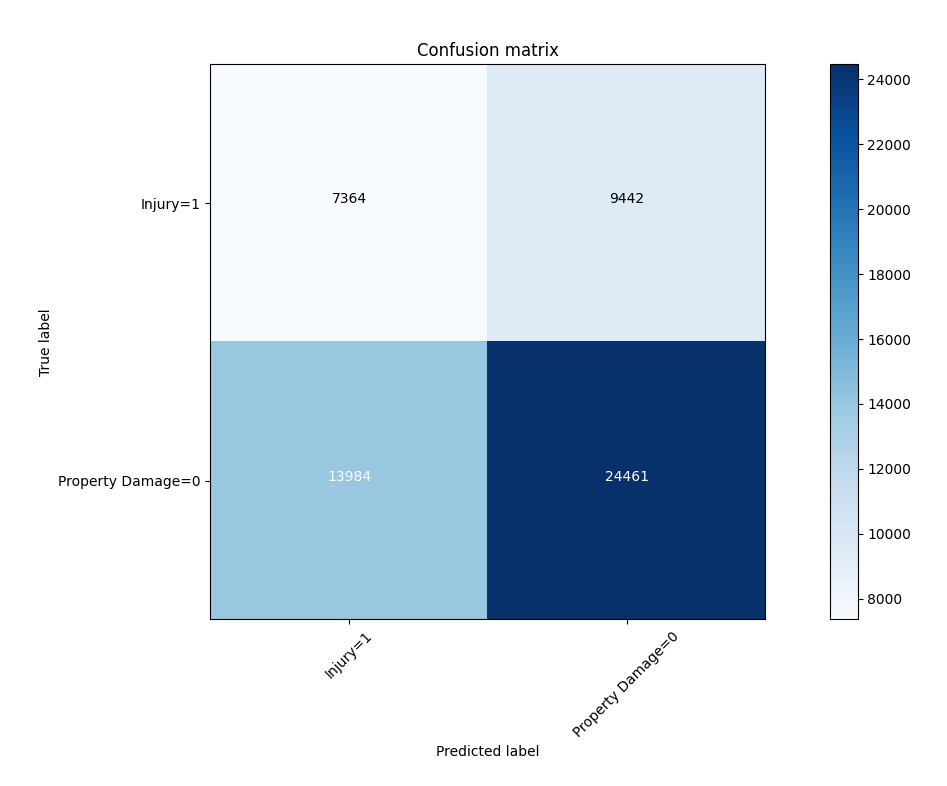
**Machine Learning Models**

Different machine learning models are used for different purposes. Logistic Regression is used to model binary dependent variables, The Decision Tree is used break down data set into smaller subset while at the same time tree is developed gradually. K-NN is used for the similarity measurement among the variables.

**Results:**

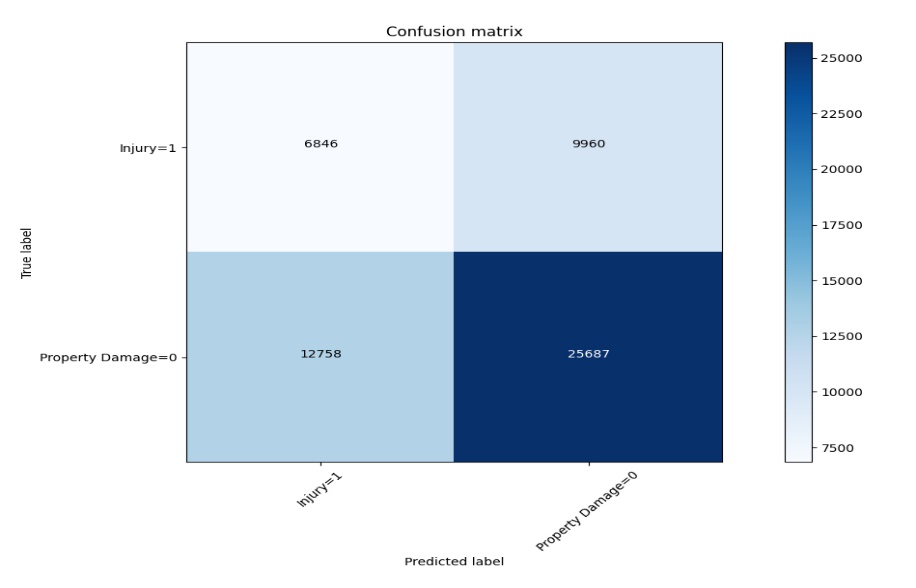
**Decision Tree Model:** The classifier used entropy and maximum depth of 6. The Evaluation Score and confusion matrix are as follow.





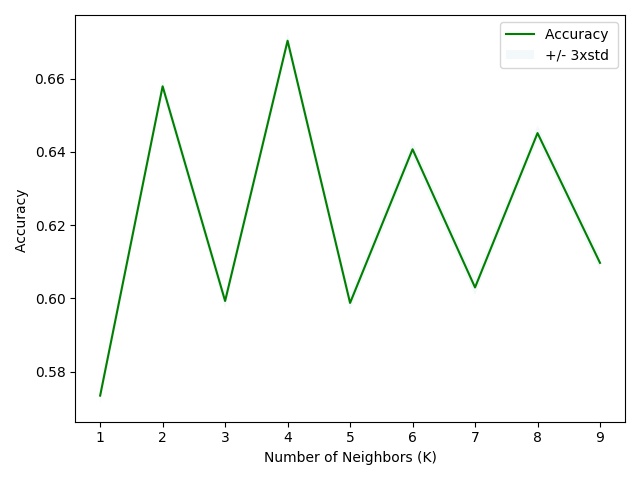
**Logistic Regression Model:** The C used for regularization strength was 0.01 whereas the function used was liblinear. The Evaluation Score and confusion matrix are as follow.

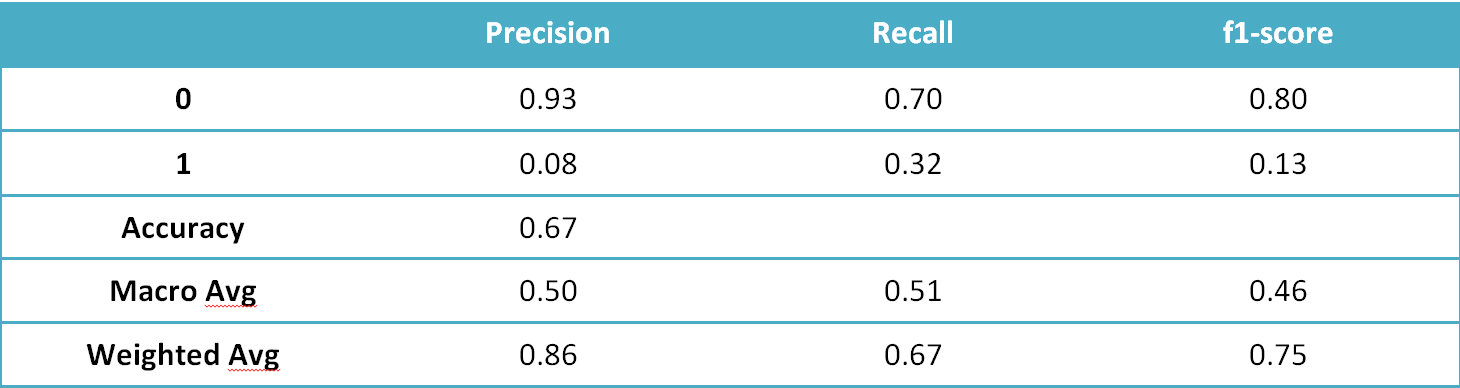




**K-Nearest Neighbor :**

The K has been choose 4.The Evaluation Score are as follow





**Conclusion**

1-The F1 score is highest for K-NN at 0.75 however it performs poor on Precision and Recall.

2-The Decision Tree Model having 0.56 F1 score and balanced precision and recall for 0 and 1.

3- The Logistic Regression F1 score 0.6 and more balanced precision and recall for 0 and 1.

So the Decision Tree and Logistics Regression shows better performance and can be used for the purpose of best fit model.

**Recommendations**

1. Road or light conditions were not ideal, if these main two factors can be improved the accidents will surely decreased.
2. The Car driver must be informed time to time about the conditions of road, lights, weather and specific locations of roads where accidents are most.