Accident Severity Prediction

Introduction / Problem:

Accidents across the world are considered as a critical and major problem. Traffic accidents can cause physical as well as financial impacts. The factors leading to accidents may involve weather conditions, speeding, road quality etc. With the help of Machine learning, we could be able to create a model to predict the severity of the accidents based on the past occurrences, which would help the drivers to be cautious and prevent accidents in the future. In this project, we would majorly consider the accidents that took place in the United States.

Data Wrangling:

The Data set used in this project is obtained from Kaggle-(<https://www.kaggle.com/sobhanmoosavi/us-accidents>) . The details available are from Feb 2016 to June 2020 with 3.5 million records. The dataset consists of various columns like Wind\_speed, weather\_condition, precipitation, pressure which would help in determining the severity of the accidents. We also have ‘sources’ as a column which specifies where the data are extracted from.

However the below listed columns are considered for this project:

1. 'Severity'
2. 'City '
3. 'County'
4. 'State'
5. 'ZipCode'
6. 'Humidity(%)'
7. 'Temperature(F)'
8. 'Pressure(in)'
9. 'Visibility(mi)'
10. 'Weather\_Condition'
11. 'Traffic\_Signal'
12. 'Junction'
13. 'Stop'

Also, only a fraction of data was used to perform the modelling as the obtained dataset is very huge. The sampled data is 1.5% (~51K records) of the total dataset. Before sampling the dataset, it was made sure that the available records have only non null values and the rest were dropped.

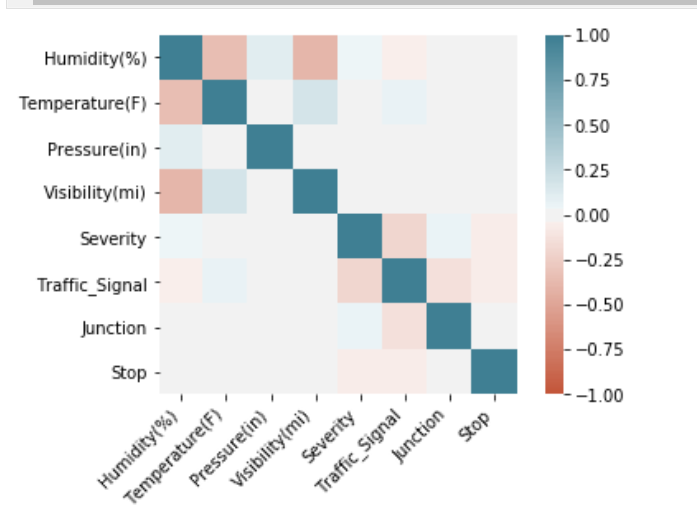
Among the considered columns only the below mentioned columns had a limited number of null values. Thus it was decided to drop all the null records.

* City 112
* Zipcode 1069
* Temperature(F) 65732
* Humidity(%) 69687
* Pressure(in) 55882
* Visibility(mi) 75856
* Weather\_Condition 76138

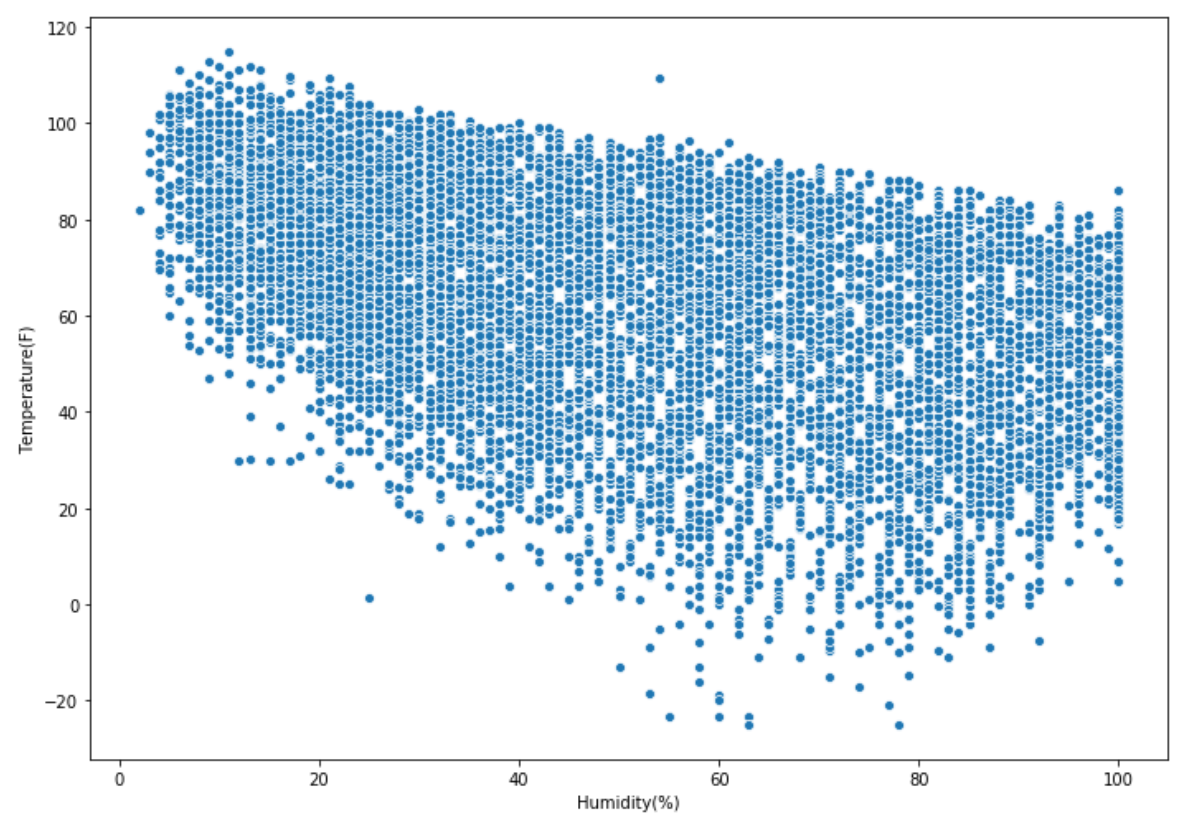
Also, the columns ‘Stop’, ‘Junction’, ‘Traffic Signal’ were converted to int from boolean as part of converting categorical variables to quantitative Variables.

Exploratory Data Analysis:

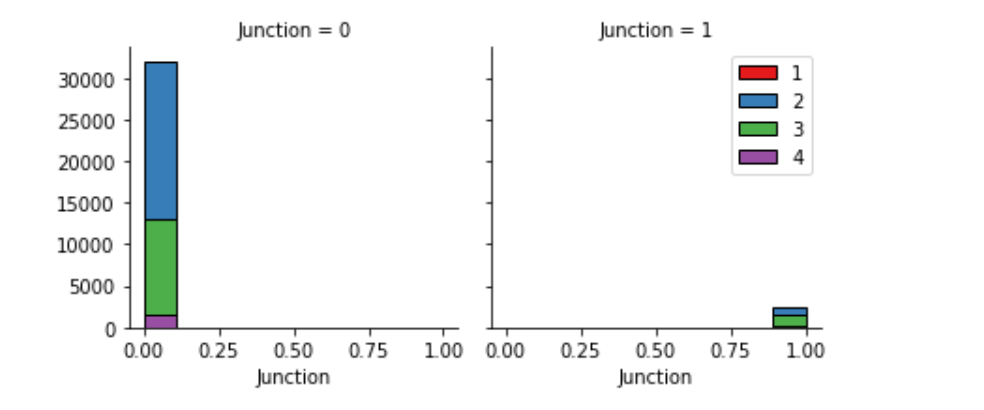
To analyze and to have a quick idea on the overall relationship between the columns, a correlation heatmap was generated as below:

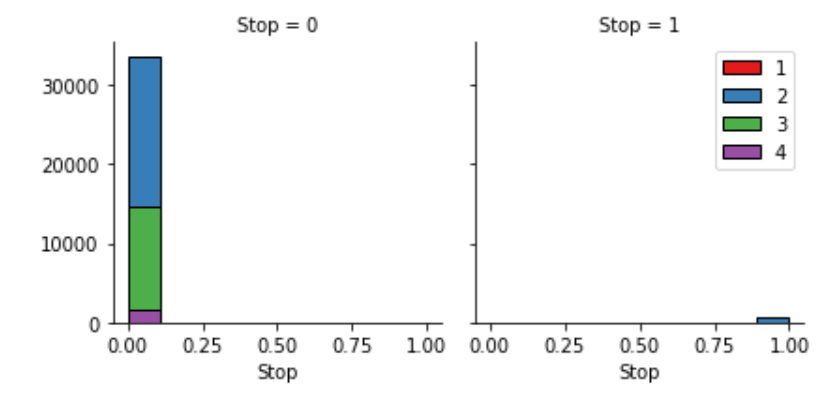


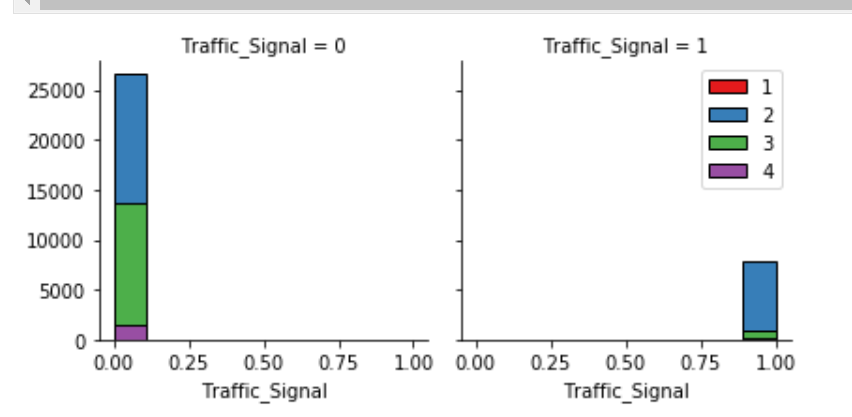
Furthermore, from the above heat map, the relation between ‘Humidity’ and ‘Temperature’ seem to have a decent -ve correlation, hence to view their relation in specific, the below scatter plot was designed.



To understand the severity of the accidents at Junctions, Stop, Traffic Signal, the below hist was created and it is clear that the number of accidents are high when there are no Junctions, Stops and Traffic Signals in nearby locations.







Modelling:

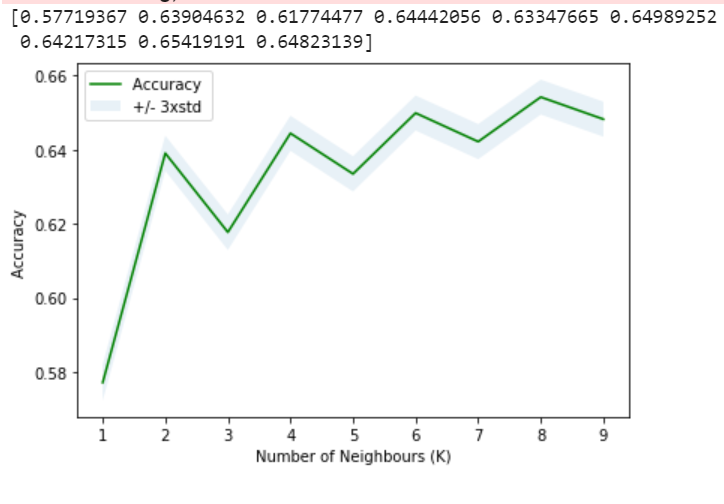
In this project I had chosen the KNN algorithm among the others as this model required no training and the dataset used was only a small fraction(~51k). The features( X ) are as follows:

* 'Humidity(%)',
* 'Temperature(F)'
* 'Pressure(in)'
* 'Visibility(mi)'
* 'Traffic\_Signal'
* 'Junction'
* 'Stop'

In addition, ‘Wind\_Conditions’ was amended by creating dummies as the values are non integers.

Target ( y ) = ‘Severity’’

Initially to determine the best K value for a better accuracy, the data was split into train & test, then the accuracy was determined with Jaccard\_Similarity\_Score for K = 0 to K=10.



The highest accuracy was noticed when **K=8**. Thus the **severity is predicted** when K=8

Results / Discussion:

The overall accuracy after prediction was 66 %, which seemed to be a decent outcome for the fraction of dataset chosen, however it can still be increased by:

* Analysing more data & column relationships.
* Considering more columns like Wind direction etc
* Performing a load test across different algorithms and with different evaluation metrics.

Conclusion:

In a nutshell, the data was analysed with correlations, histograms which provided an idea to determine the frequent cause of accidents.

Considering the above analysis, the dataset fraction and the columns chosen, KNN modelling was performed with the best K value and the severity is predicted successfully.