USA - Road Accident Traffic Delay Severity Prediction

Tholkappiyan Aranganathan
Oct 03, 2020

www.linkedin.com/in/tholkappiyan-aranganathan-287b1021 | a_tholkappiyan@yahoo.com

Introduction

Road accident become very common these days and it is increasing gradually as the infrastructure and technology improves. The main causes for road accident are high speed, speaking on cell phone while driving, drunk and driving, weather condition and many more.

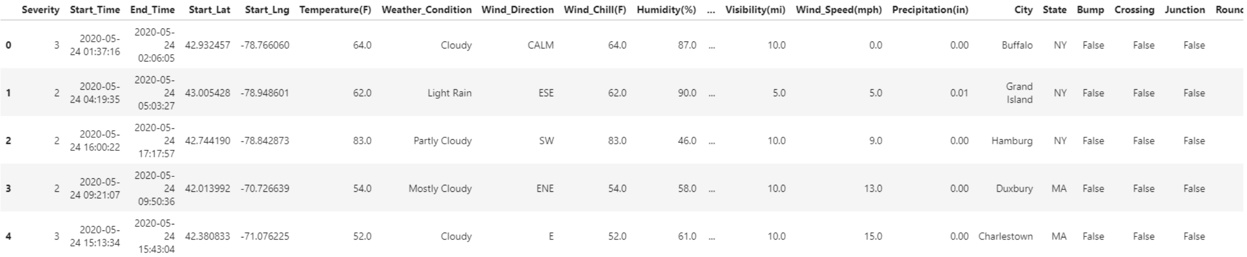
Approximately 1.3 million people die each year in the road accident across the countries. Road traffic injuries are the leading cause of death for children and youngster aged between 5 – 29. Another fact is that in the road accident more than 70% of deaths occurs among young male which put their entire family's future at risk if he is the only earning person in the family. It will also affect the children education and economy.

Problem

This project aims to predict the road accident traffic delay severity. Predicting the traffic delay severity will help to avoid traffic jam at peak hours, based on the delay severity the traffic can be diverted to an alternate route to reach the destination, it will ensure the essential services like medical service, food delivery, fire service, and so on can reach on time to the destination.

Data Collection

I have downloaded the data from https://www.kaggle.com/. The dataset is having 49 features with a sample of 2 million records. The dataset is having a combination of nominal, ordinal and numerical data.



	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	Temperature(F)	Weather_Condition	Wind_Direction	Wind_Chill(F)	Humidity(%)	...	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)	City	State	Bump	Crossing	Junction	Rounc
0	3	2020-05-24 01:37:16	2020-05-24 02:06:05	42.932457	-78.766060	64.0	Cloudy	CALM	64.0	87.0	...	10.0	0.0	0.00	Buffalo	NY	False	False	False	
1	2	2020-05-24 04:19:35	2020-05-24 05:03:27	43.005428	-78.948601	62.0	Light Rain	ESE	62.0	90.0	...	5.0	5.0	0.01	Grand Island	NY	False	False	False	
2	2	2020-05-24 16:00:22	2020-05-24 17:17:57	42.744190	-78.842873	83.0	Partly Cloudy	SW	83.0	46.0	...	10.0	9.0	0.00	Hamburg	NY	False	False	False	
3	2	2020-05-24 09:21:07	2020-05-24 09:50:36	42.013992	-70.726639	54.0	Mostly Cloudy	ENE	54.0	58.0	...	10.0	13.0	0.00	Duxbury	MA	False	False	False	
4	3	2020-05-24 15:13:34	2020-05-24 15:43:04	42.380833	-71.076225	52.0	Cloudy	E	52.0	61.0	...	10.0	15.0	0.00	Charlestown	MA	False	False	False	

Data Preprocessing

Data Cleaning - Some of the columns are having null records and duplicate values. I have replaced the null records with mean and mode based on the data type. The accident time and duration has derived from the accident start_time and end_time. The columns "Weather Condition" and "Wind Direction" are categorical variables. As the machine learning model will not work with categorical data, I have converted these variables into binary values by using one hot encoding. Also there few features are in Boolean format, those values are converted to numeric value by changing its datatype to integer.

Methodology

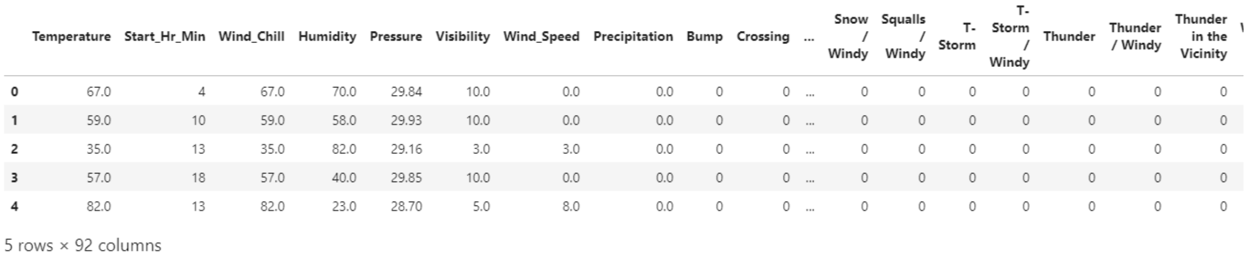
The target attribute is a categorical variable. Hence, we will use Classification algorithm to build the model. Classification is a supervised learning approach. Used to categorizing some unknown items into a discrete set of categories or classes. The machine learning algorithm model will be developed using the popular classification algorithms K-Nearest Neighbors, Logistic Regression, Decision Tree and Support Vector Machine. The accuracy of the all the model will be evaluated using different metrics like Jaccard Score, F1 Score and Logloss. The model with high accuracy will be deployed in the production.

Feature Selection

After the data wrangling there are 49 features with 500609 samples. Out of which I have selected the below features which are having more impact on the predict variable traffic delay severity

- Temperature
- Start_Hr_Min
- Wind_Chill
- Humidity

- Pressure
- Visibility
- Wind_Speed
- Wind_Direction
- Weather_Condition
- Precipitation
- Bump
- Crossing
- Junction
- Roundabout
- Stop



	Temperature	Start_Hr_Min	Wind_Chill	Humidity	Pressure	Visibility	Wind_Speed	Precipitation	Bump	Crossing	...	Snow / Windy	Squalls / Windy	T-Storm	T-Storm / Windy	Thunder	Thunder / Windy	Thunder in the Vicinity
0	67.0	4	67.0	70.0	29.84	10.0	0.0	0.0	0	0	...	0	0	0	0	0	0	0
1	59.0	10	59.0	58.0	29.93	10.0	0.0	0.0	0	0	...	0	0	0	0	0	0	0
2	35.0	13	35.0	82.0	29.16	3.0	3.0	0.0	0	0	...	0	0	0	0	0	0	0
3	57.0	18	57.0	40.0	29.85	10.0	0.0	0.0	0	0	...	0	0	0	0	0	0	0
4	82.0	13	82.0	23.0	28.70	5.0	8.0	0.0	0	0	...	0	0	0	0	0	0	0

5 rows × 92 columns

Exploratory Analysis

As a first step, we will visualize the places which are recorded more than 250 accidents, I am using the Pandas Folium library to draw this map. Folium map requires the latitude and longitude to spot the place. This map provides options to zoom in by double clicking on it. I have added a tool tip message to show the place and accident severity details, by click on the circle mark it will display the State, City, Severity and total number of accidents occurred at the selected place.

All the features of matrix are self-explanatory. However, it is important to understand the meaning of predicting vector Severity.

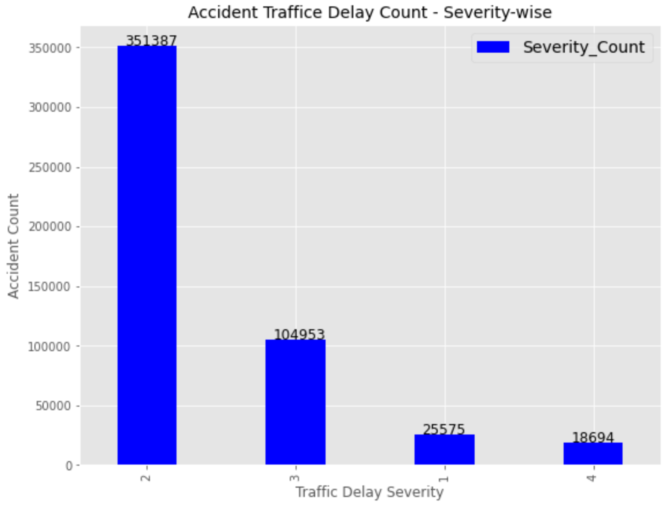
Severity – 1: An accident with less traffic delay within 1 hour

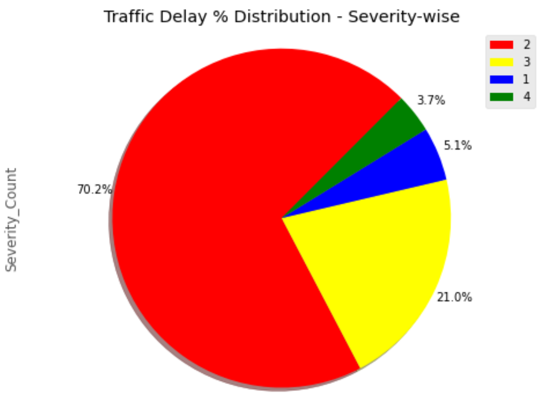
Severity – 2: An accident with traffic delay between 1 hours to 3 hours

Severity – 3: An accident with traffic delay between 3 hours to 6 hours

Severity – 4: An accident with traffic delay more than 6 hours

Let us visualize the data for a better understanding which is very critical to select more relevant features for the machine learning model.

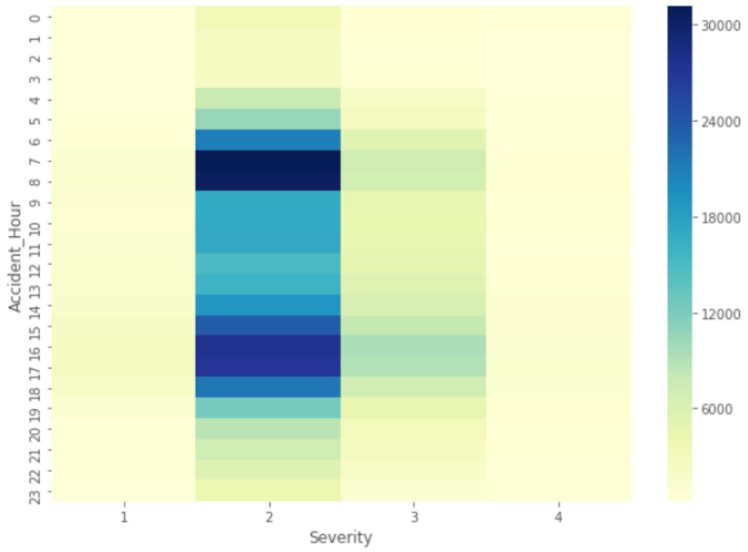




The bar chart depicts that majority of the accident's traffic delay impact is with severity 2, following to that severity 3, 1 and 4. The pie chart gives the % distribution of the traffic delay severity. It is evident that 70.2 % of accidents are occurred with severity 2. The severity 3 & 4 together is 24.7%.

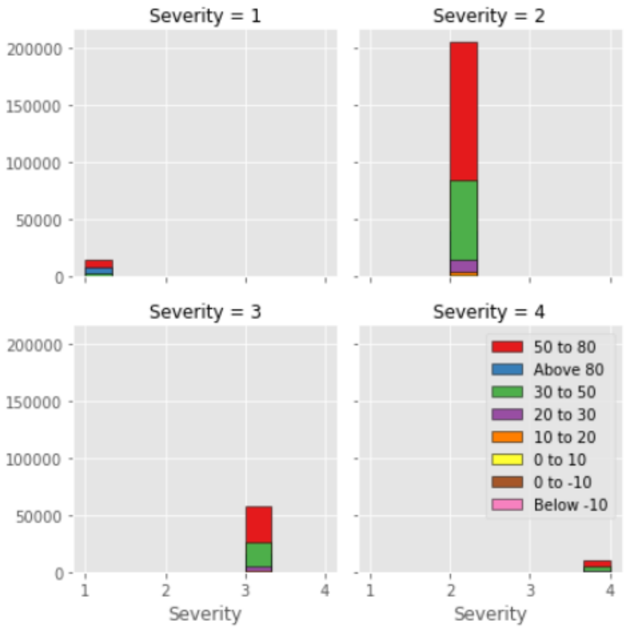
Let us understand the state wise accident count. The below picture shows that a huge number of accidents are recorded in the California state. It is more than 5 time compared to the other busiest and highly populated states like South Carolina, Florida, Texas and so on.

Relation Between Accident Time and Severity



From the heatmap, it is very clear that majority of high severity traffic delay accidents were occurred in morning between 6.00 AM to 8.00 AM and, in the afternoon & evening between 2.00 PM to 6.00 PM

Relationship Between Temperature and Severity

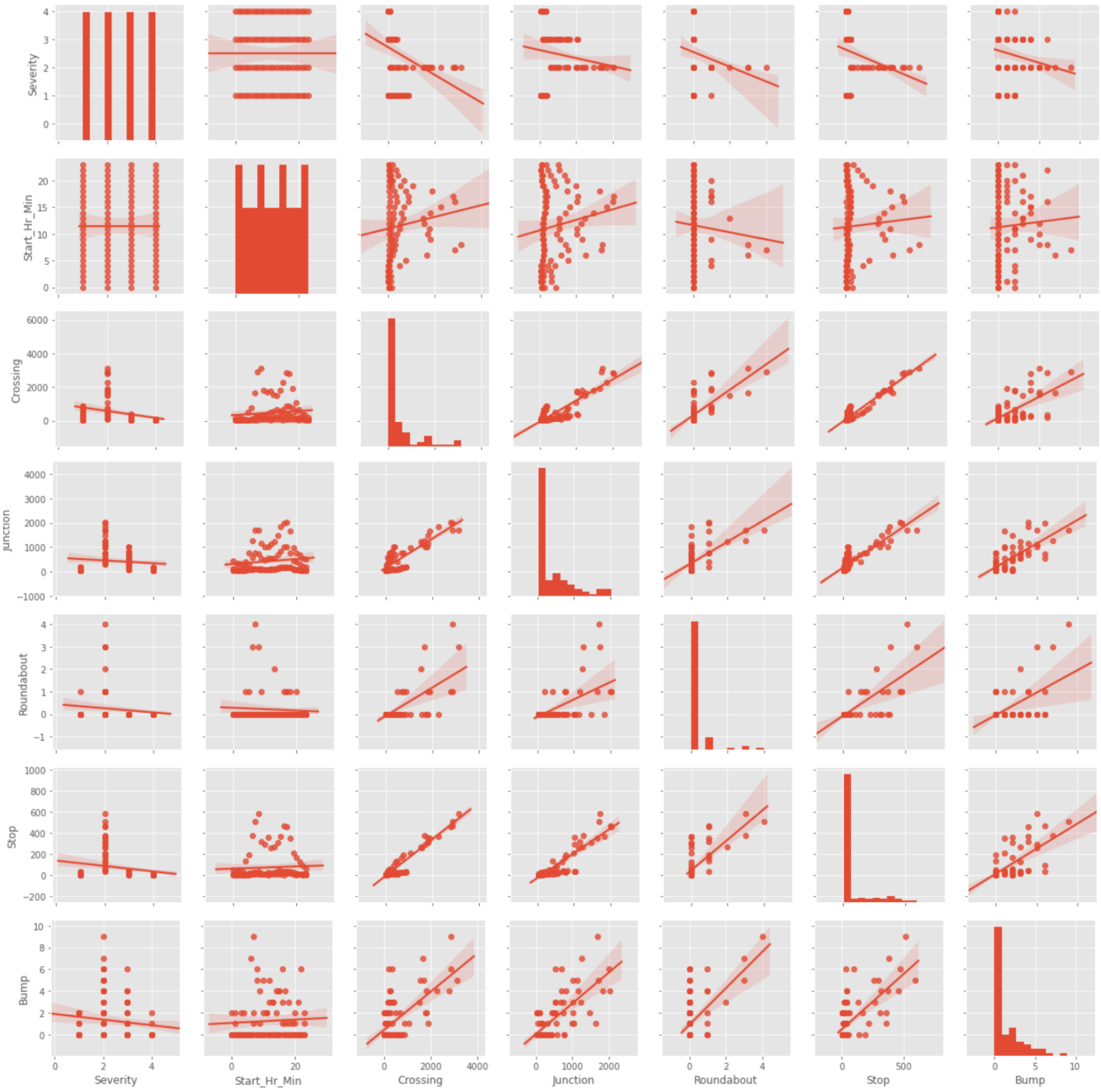


From the chart we can understand that a greater number of accidents happened in the temperature between 50 - 80-degree Fahrenheit and above. We may conclude that the accident severity is not having any correlation with the temperature.

There is no correlation between the traffic severity and visibility. All the accidents are happened only with the visibility greater than 10 miles.

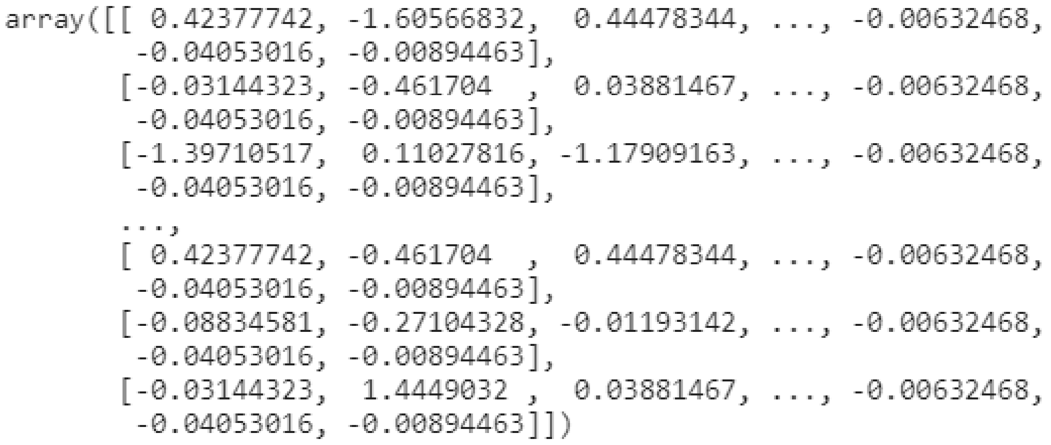
Relationship between Bump, Junction, Crossing, Roundabout with Severity

The pair chart is self-explanatory. Some of the matrix features like Junction, Crossing, Roundabout, Stop and Bump are having linear relationship with the Y vector.



Feature Scaling

The matrix of features has been normalized before building the model using sklearn StandardScalar module.



```
array([[ 0.42377742, -1.60566832,  0.44478344, ..., -0.00632468,
        -0.04053016, -0.00894463],
       [-0.03144323, -0.461704  ,  0.03881467, ..., -0.00632468,
        -0.04053016, -0.00894463],
       [-1.39710517,  0.11027816, -1.17909163, ..., -0.00632468,
        -0.04053016, -0.00894463],
       ...,
       [ 0.42377742, -0.461704  ,  0.44478344, ..., -0.00632468,
        -0.04053016, -0.00894463],
       [-0.08834581, -0.27104328, -0.01193142, ..., -0.00632468,
        -0.04053016, -0.00894463],
       [-0.03144323,  1.4449032 ,  0.03881467, ..., -0.00632468,
        -0.04053016, -0.00894463]])
```

Train Test Split

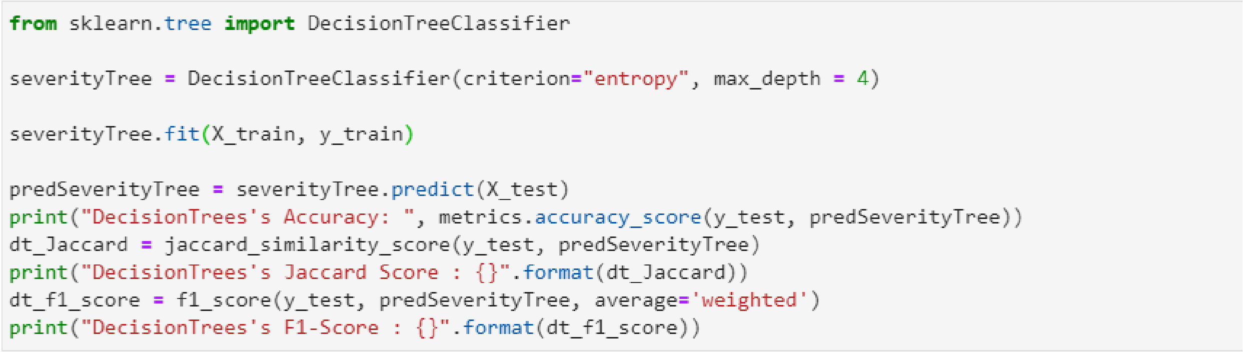
The dataset has been split for train and test the model. I have used the train_test_split module from sklearn.model_selection class. The dataset has been split with 80:20 ration for train and test the model respectively. I have chosen a random state of 4.

Predictive Model

I have developed two model for this project. Decision Tree and Logistic Regression and compared both models' metrics using Jaccard Score and F1-Score. The details are given below for reference.

Decision Tree

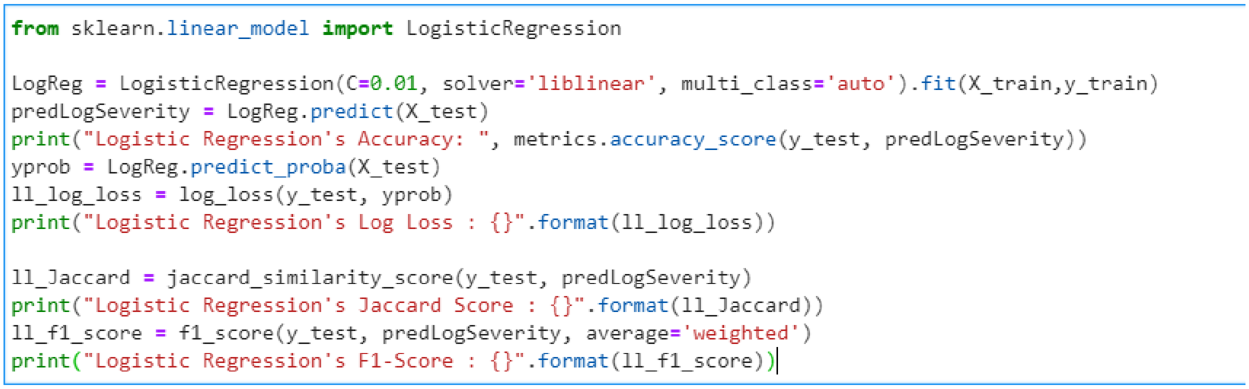
I have imported the DecisionTreeClassifier from sklearn tree library, I have used the criterion as entropy and the max_depth of the decision tree to 4. Because the accident severity is classified into four groups. The model has been trained by calling the fit method of the DecisionTreeClassifier and passed the X_train and y_train dataset. The model accuracy has tested by using X_test data and compared using Jaccard Score and F1-Score



```
from sklearn.tree import DecisionTreeClassifier

severityTree = DecisionTreeClassifier(criterion="entropy", max_depth = 4)

severityTree.fit(X_train, y_train)

predSeverityTree = severityTree.predict(X_test)
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, predSeverityTree))
dt_Jaccard = jaccard_similarity_score(y_test, predSeverityTree)
print("DecisionTrees's Jaccard Score : {}".format(dt_Jaccard))
dt_f1_score = f1_score(y_test, predSeverityTree, average='weighted')
print("DecisionTrees's F1-Score : {}".format(dt_f1_score))
```

Logistic Regression

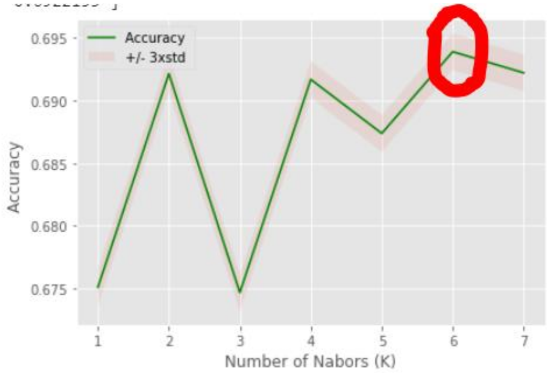
Logistic Regression model is developed with same dataset as used in the decision tree, this basically to compare the which model suits best for the given dataset. The Logistic Model's accuracy has tested using the log loss metrics, Jaccard Score and F1-Score. The comparison between these two models has given in Model Evaluation Section.

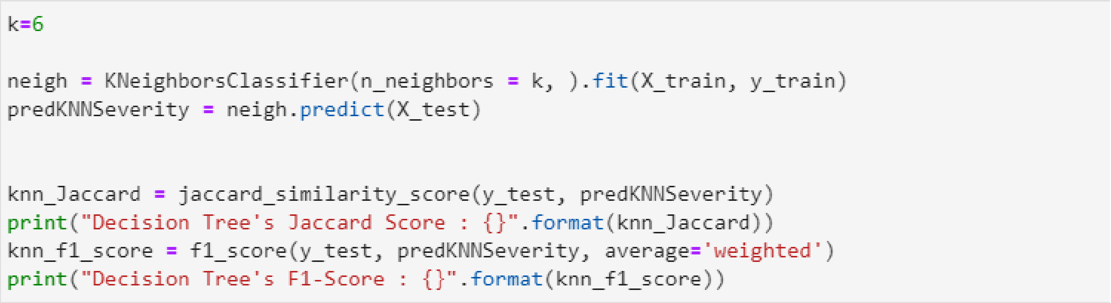


```
from sklearn.linear_model import LogisticRegression

LogReg = LogisticRegression(C=0.01, solver='liblinear', multi_class='auto').fit(X_train,y_train)
predLogSeverity = LogReg.predict(X_test)
print("Logistic Regression's Accuracy: ", metrics.accuracy_score(y_test, predLogSeverity))
yprob = LogReg.predict_proba(X_test)
ll_log_loss = log_loss(y_test, yprob)
print("Logistic Regression's Log Loss : {}".format(ll_log_loss))

ll_Jaccard = jaccard_similarity_score(y_test, predLogSeverity)
print("Logistic Regression's Jaccard Score : {}".format(ll_Jaccard))
ll_f1_score = f1_score(y_test, predLogSeverity, average='weighted')
print("Logistic Regression's F1-Score : {}".format(ll_f1_score))
```

K-Nearest Neighbor

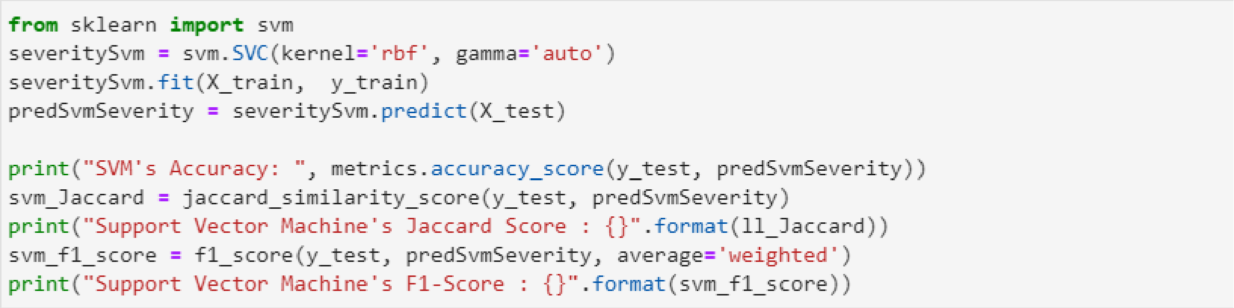
K-Nearest Neighbor model also developed with same dataset. The optimum value for K has found using the Elbow method.



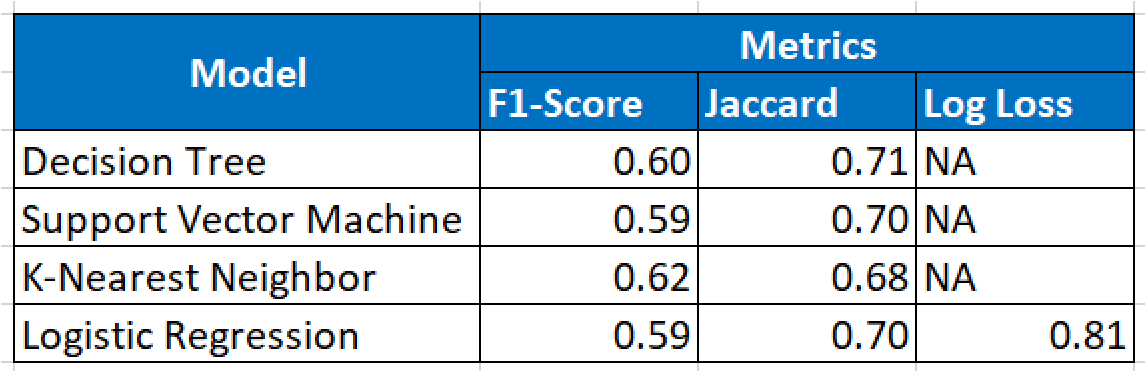

```
k=6

neigh = KNeighborsClassifier(n_neighbors = k, ).fit(X_train, y_train)
predKNNSeverity = neigh.predict(X_test)


knn_Jaccard = jaccard_similarity_score(y_test, predKNNSeverity)
print("Decision Tree's Jaccard Score : {}".format(knn_Jaccard))
knn_f1_score = f1_score(y_test, predKNNSeverity, average='weighted')
print("Decision Tree's F1-Score : {}".format(knn_f1_score))
```

Support Vector Machine

The last model developed for this project is using the support vector machine. I have taken the kernel as ‘rbf’ and gamma as auto get the best output for this dataset.


```
from sklearn import svm
severitySvm = svm.SVC(kernel='rbf', gamma='auto')
severitySvm.fit(X_train,  y_train)
predSvmSeverity = severitySvm.predict(X_test)

print("SVM's Accuracy: ", metrics.accuracy_score(y_test, predSvmSeverity))
svm_Jaccard = jaccard_similarity_score(y_test, predSvmSeverity)
print("Support Vector Machine's Jaccard Score : {}".format(ll_Jaccard))
svm_f1_score = f1_score(y_test, predSvmSeverity, average='weighted')
print("Support Vector Machine's F1-Score : {}".format(svm_f1_score))
```

Model Evaluation



Model	Metrics		
	F1-Score	Jaccard	Log Loss
Decision Tree	0.60	0.71	NA
Support Vector Machine	0.59	0.70	NA
K-Nearest Neighbor	0.62	0.68	NA
Logistic Regression	0.59	0.70	0.81

Solution to the Problem

From the exploratory data analysis, it has been concluded that the high severity accidents are happening

1. In the morning time between 6.00 AM and 8.00 AM and 3.00 PM to 6.00 PM
2. Accidents near Bump, Junction, Roundabout, Crossing and Stop are recorded high severity traffic delay

The traffic controller can take preventive measures in the above set area to avoid the accidents and reduce the high severity traffic delay.