PREDICTION THE SEVERITY OF AN ACCIDENT (CASE STUDY)

KONDA LAVANYA September 22,2020

Introduction

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.

It would be great if real-time conditions can be provided to estimate the trip safeness. In this way, it can be decided beforehand if the driver will take the risk, based on reliable information.

Business Understanding

Predicting crash injury severity is a crucial constituent of reducing the consequences of traffic crashes.The number of traffic crashes and their victims has been a rising trend globally due to increases in population and motorization. Different factors involved in traffic crashes have a substantial effect on each other, thus making it difficult to individually consider any of the parameters when explaining the severity of traffic crashes.

We can use the Machine learning models to Predict the severity of the Accidents based on the factors like the type of Road travelling,Location,number of Vehicles,Number of People on Road,weather any many more. These models Helps Road Users, Insurance Companies,Health Care providers,Government etc.

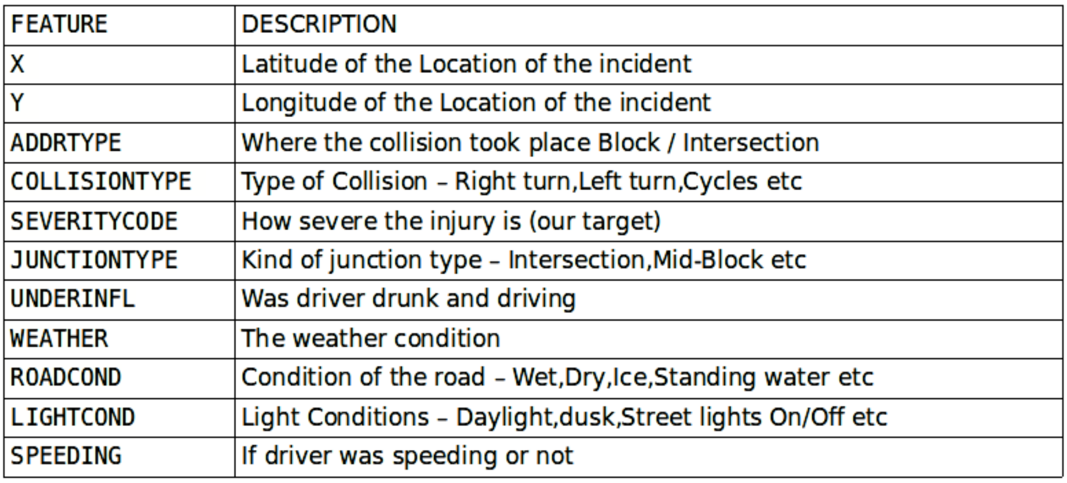
This help user to have a better understanding of Road Conditions, their impacts, helps to take initiatives to reduce the Accidents and to build New Infrastructure.

Data Set

I worked on Seattle GeoData to Predict Accident Severity in Seattle.provided by the Traffic Records Group in the SDOT Traffic Management Division from Seattle, WA. It includes all collisions provided by the Seattle Police Department and recorded by the Traffic Record, displayed at the intersection or mid-block of a segment from 2004 to the present.

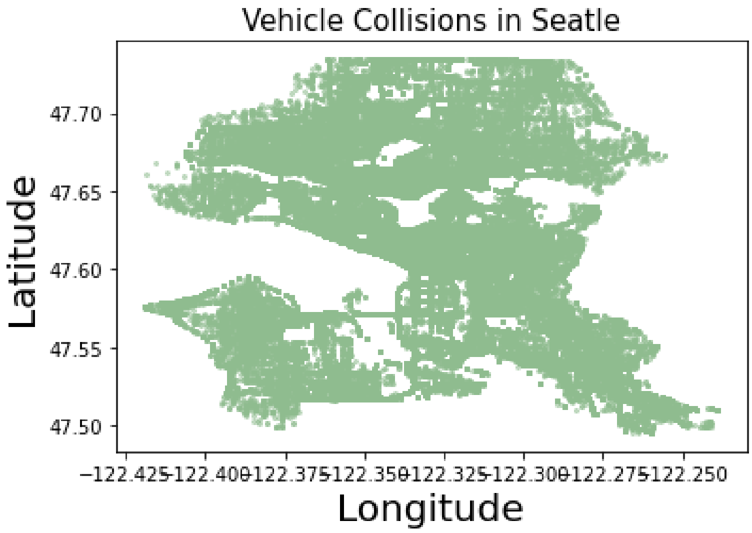
Data understanding

The Dataset Consists of 40 features which are X,Y,OBJECTID,INCKEY,COLDETKEY,REPORTNO,STATUS,ADDRTYPE,INTKEY,LOCATION, EXCEPTRSNCODE,EXCEPTRSNDESC,SEVERITYCODE,SEVERITYDESC,COLLISIONTYPE, PERSONCOUNT,PEDCOUNT,PEDCYLCOUNT,VEHCOUNT,INJURIES,SERIOUSINJURIES, FATALITIES,INCDATE,INCDTTM,JUNCTIONTYPE,SDOT_COLCODE,SDOT_COLDESC, INATTENTIONIND,UNDERINFL,WEATHER,ROADCOND,LIGHTCOND,PEDROWNOTGRNT, SDOTCOLNUM,SPEEDING,ST_COLCODE,ST_COLDESC,SEGLANEKEY,CROSSWALKKEY, HITPARKEDCAR
Description of few features is as below



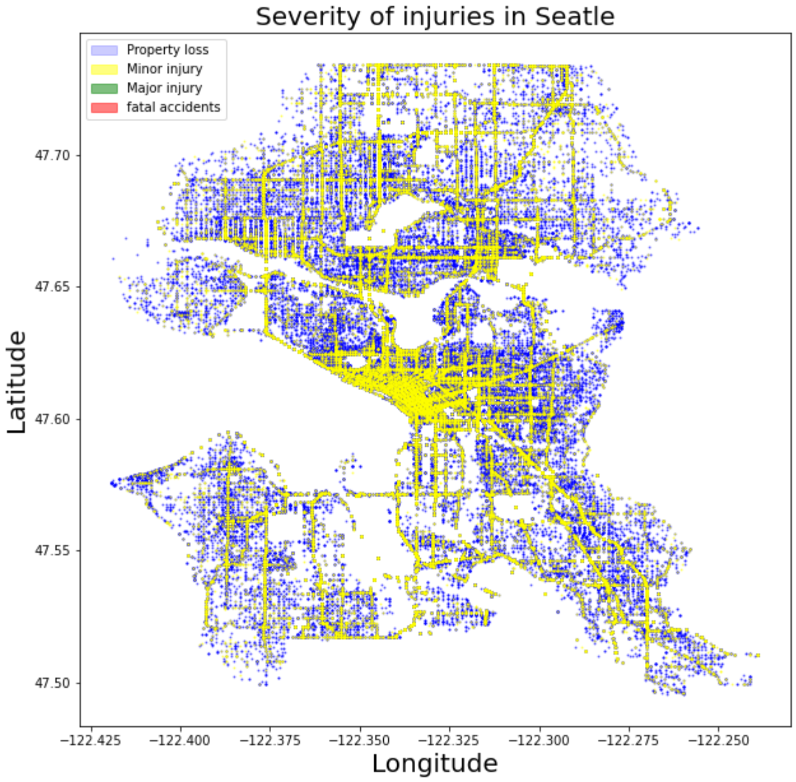
FEATURE	DESCRIPTION
X	Latitude of the Location of the incident
Y	Longitude of the Location of the incident
ADDRTYPE	Where the collision took place Block / Intersection
COLLISIONTYPE	Type of Collision – Right turn,Left turn,Cycles etc
SEVERITYCODE	How severe the injury is (our target)
JUNCTIONTYPE	Kind of junction type – Intersection,Mid-Block etc
UNDERINFL	Was driver drunk and driving
WEATHER	The weather condition
ROADCOND	Condition of the road – Wet,Dry,Ice,Standing water etc
LIGHTCOND	Light Conditions – Daylight,dusk,Street lights On/Off etc
SPEEDING	If driver was speeding or not

Lets us plot and try to understand the Vehicle Collisions in Seattle city.



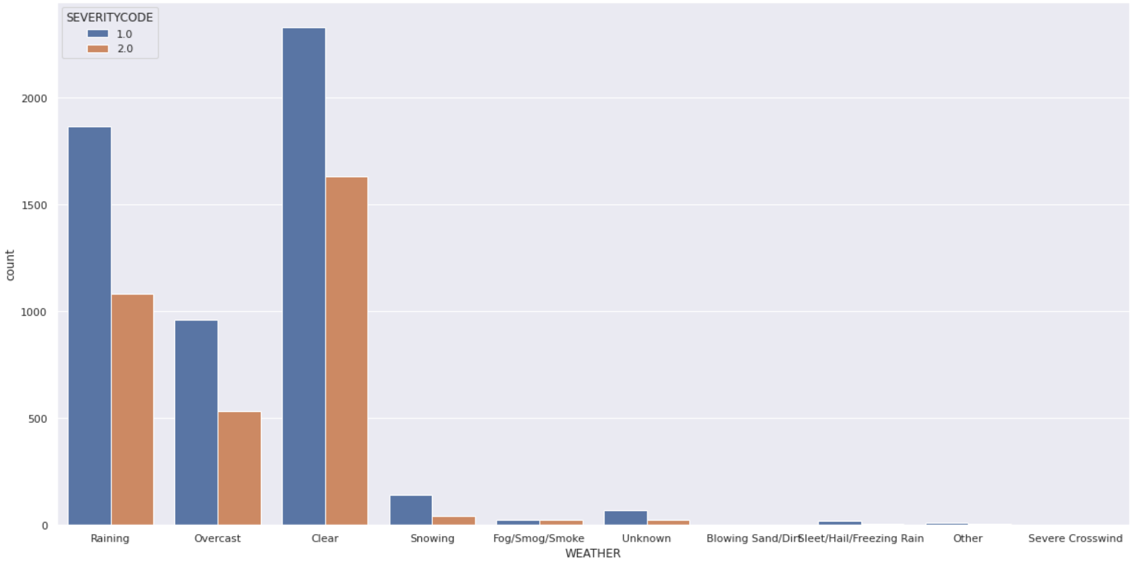
I have Plotted all the X,Y i.e Latitude and Longitude of places where the incident has taken place using scatter Plot. The plot indicates that the accidents are taking places All over Seattle. now lets try to plot the severity of the collisions on a Scatter plot.

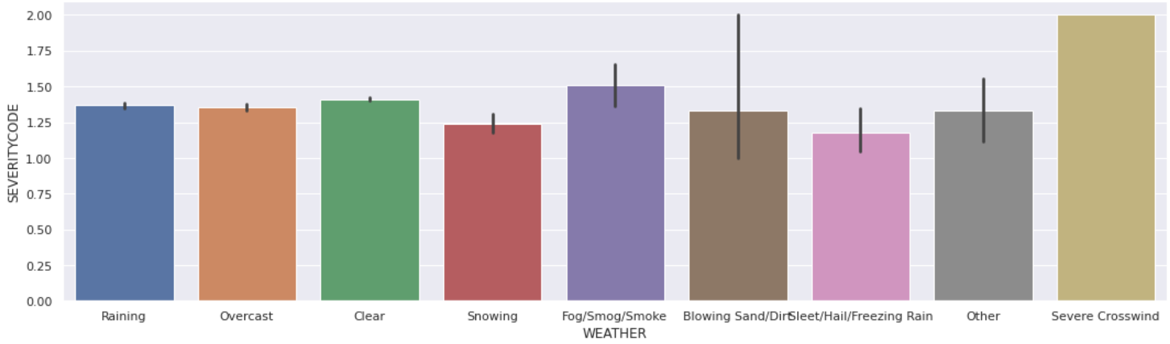
- 'blue' -'Property loss'
- 'yellow' -'Minor injury'
- 'green' - 'Major injury
- 'red' - 'Accident'



Data Exploration:

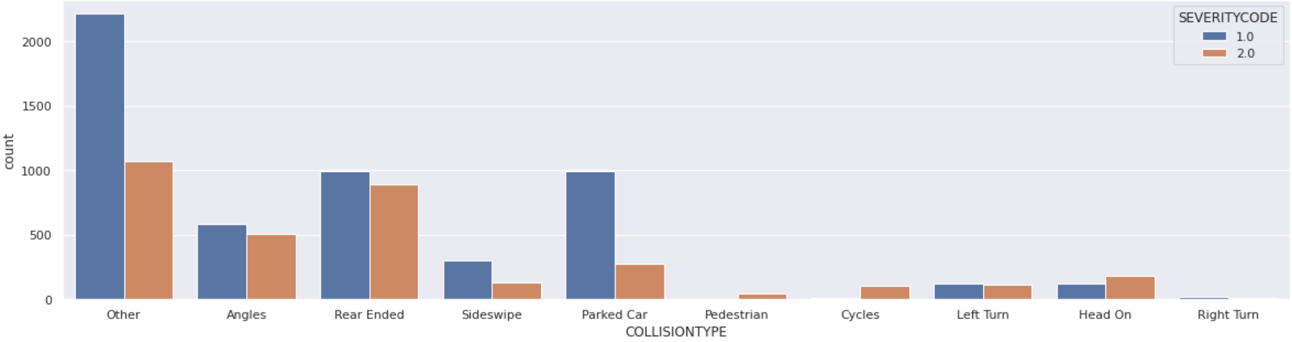
- To Understand the Features better I have plotted all the Features vs Severity code.
- **Weather vs SEVERITYCODE** is as shown Below, From the plots below Collisons effect more in Severe Crosswind





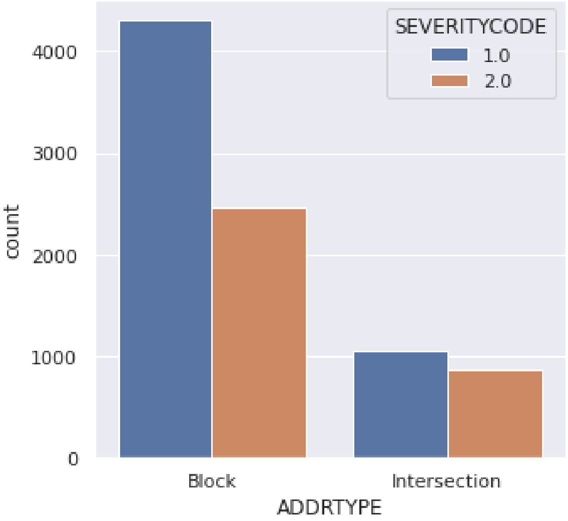
Collision type vs SEVERITYCODE

- COLLISIONTYPE vs SEVERITYCODE is as shown Below, From the plots below Rear Ended,Parked Car has more effect on Severity of Collision.



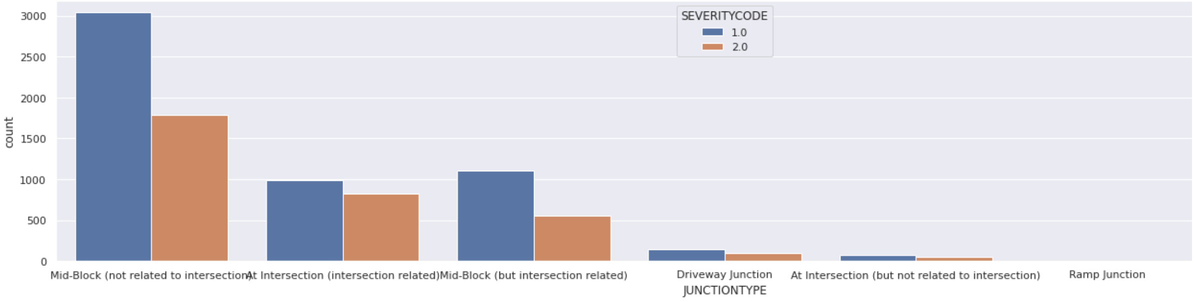
Collision place vs SEVERITYCODE

- COLLISIONTYPE vs SEVERITYCODE is as shown Below, From the plots below Block Area has more effect on Severity of Collision.



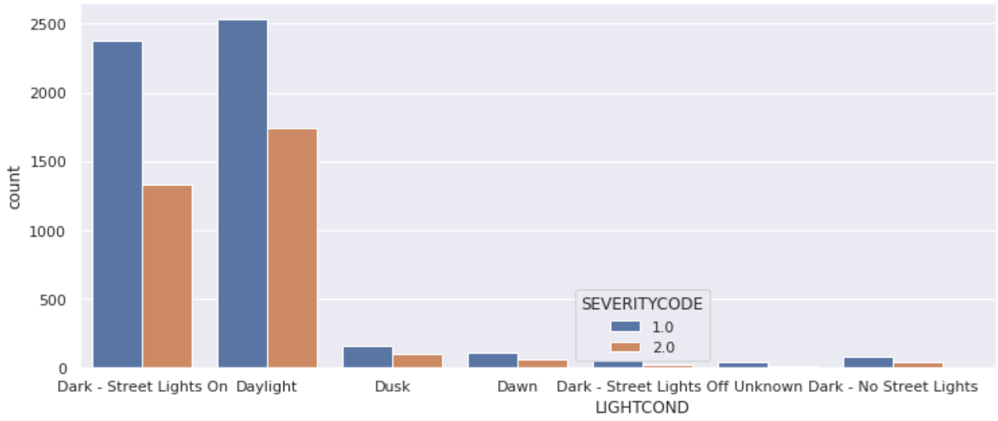
Junction type vs SEVERITYCODE

- JUNCTIONTYPE vs SEVERITYCODE is as shown Below, From the plots below Block Area has more effect on Severity of Collision.



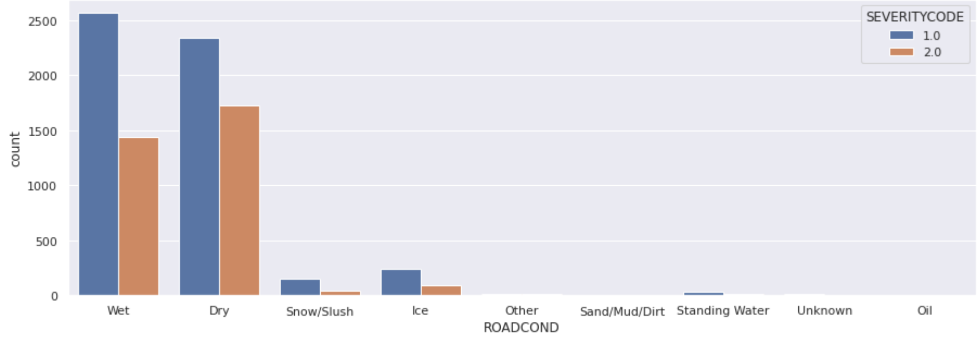
Light condition vs SEVERITYCODE

- LIGHTCOND vs SEVERITYCODE is as shown Below, From the plots below Collisions take place in daytime or when the light is avaliable.



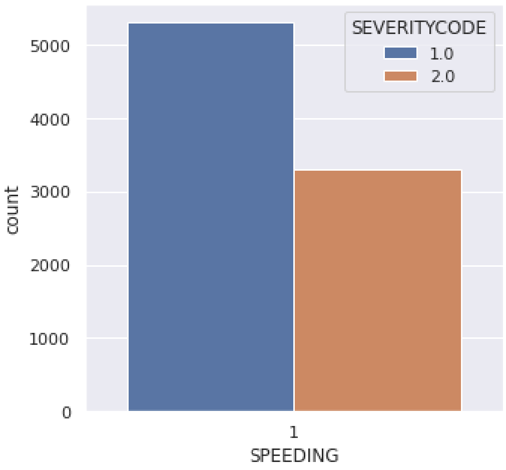
Road condition vs SEVERITYCODE

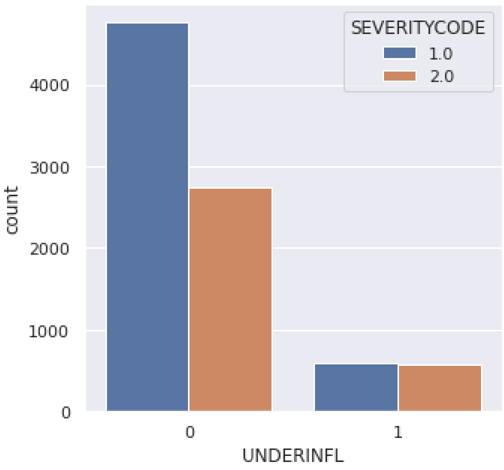
- ROADCOND vs SEVERITYCODE is as shown Below, From the plots below Collisions take place in wet and Dry Roads.



Speeding vs SEVERITYCODE / UNDERINFL vs SEVERITYCODE

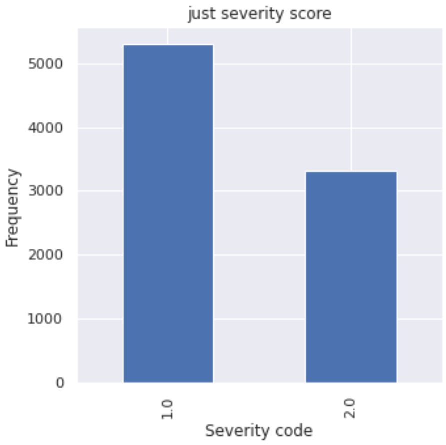
- Speeding vs SEVERITYCODE / UNDERINFL vs SEVERITYCODE are as shown Below, From the plots below Collisions many collisions take place even when driver is not drunk

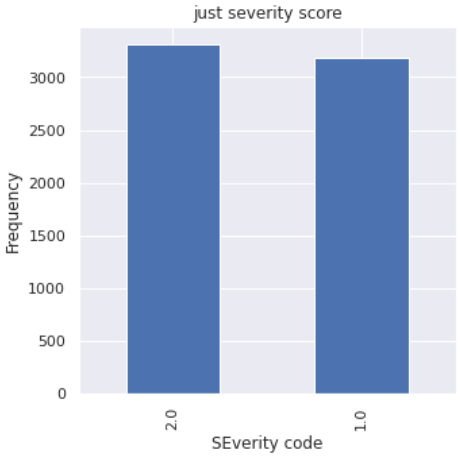




Data Preparation:

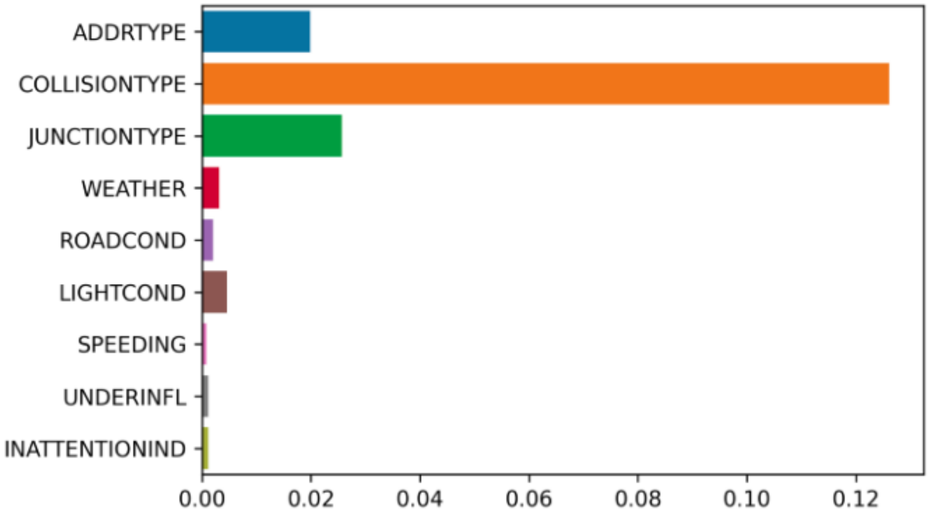
- Droping all the rows with values Nan / Unknown
- Changing the Values Y to 1 and N to 0
- Since the property Loss and Minor injuries are common So lets drop the rows with SEVERITYCODE - 0,2b,3.
- After checking the Number of rows with SEVERITYCODE 1,2. As we are required to make a balanced dataset we are dropping a few random rows with SEVERITYCODE 1
- the plots show the counts before and after balancing data.





Feature Selection:

One of the most important questions before training the model is, are all the features adding the same information to the model? If not so, what variables have more weight on it? To tackle this question some techniques can be used to help select the important features, the ones adding more information to our model. It has to be taken into mind that categorical inputs and output will be used, hence, for this kind of variables I used Mutual Information Feature Selection.Finally I choosed to work with "ADDRTYPE","COLLISIONTYPE","JUNCTIONTYPE","UNDERINFL","WEATHER","ROADCOND","LIGHTCOND","SPEEDING","SEVERITYCODE"

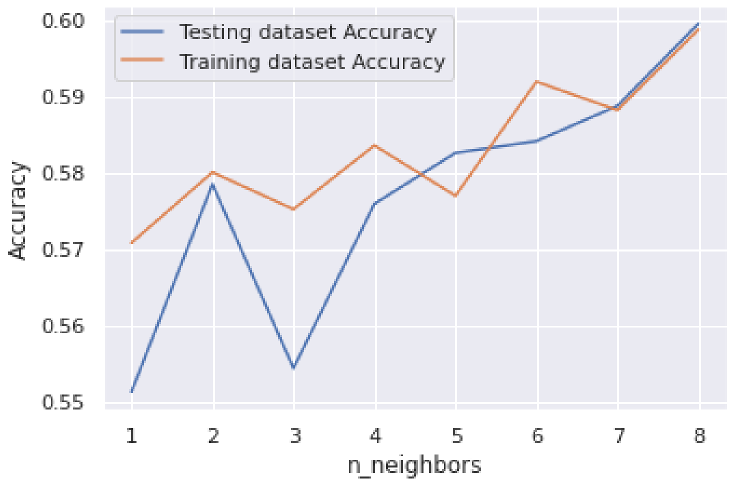


Modelling:

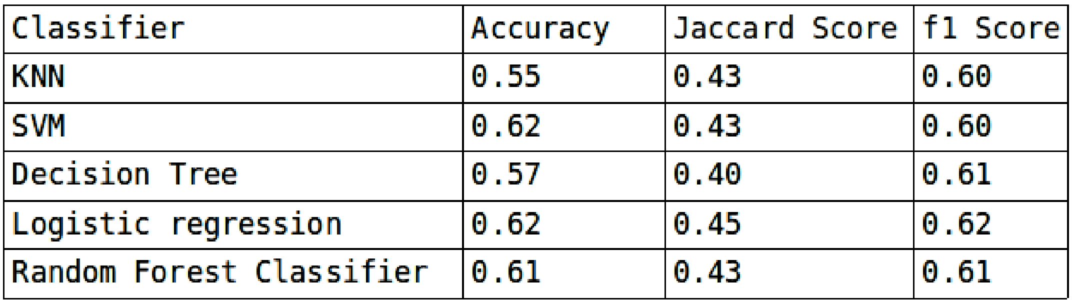
I have worked on the following Classification Methods

- KNN Classifier
- SVM
- Decision tree
- Logistic Regression
- Random forest Classifier

The first thing is to find a good value of K. The graph below shows that k=7 is good one to choose So I worked with that value for K.



Results



Classifier	Accuracy	Jaccard Score	f1 Score
KNN	0.55	0.43	0.60
SVM	0.62	0.43	0.60
Decision Tree	0.57	0.40	0.61
Logistic regression	0.62	0.45	0.62
Random Forest Classifier	0.61	0.43	0.61

Conclusions

- Collisions which does not involve personal injuries are twice as frequent as the ones involving damaged people.
- Accidents involving cycles or pedestrians are severe and involves injuries.
- The riskier car collisions are the ones that hit the car from the rear end.
- Intersection collisions are one of the most common types of crash
- Left turns are also risky maneuvers which should also be avoided if the road users want to be safe.
- Extremely dangerous weather and road conditions do not produce a quite significant accident rate, such as snow and ice. However ,caution have to be taken with rainy weather and wet roads, since after clear days and dry roads, these are the following conditions in order of importance.