Dataset :

This is an standard data set published by the Seattle Police Department, with over 194673 observations with 37 attributes collected over the last 15 years. By this huge data we have to make a effective model to prevent future accident and reduce severity, so it can be use by people for getting security and also use by companies to build a reilable system .

Attributes : SEVERITYCODE , X , Y , OBJECTID , INCKEY , COLDETKEY , REPORTNO , STATUS , ADDRTYPE , INTKEY , LOCATION , EXCEPTRSNCODE , EXCEPTRSNDESC , SEVERITYCODE.1 , SEVERITYDESC , COLLISIONTYPE , PERSONCOUNT , PEDCOUNT , PEDCYLCOUNT , VEHCOUNT , INCDATE , INCDTTM , JUNCTIONTYPE , SDOT_COLCODE , SDOT_COLDESC , INATTENTIONIND , UNDERINFL , WEATHER , ROADCOND , LIGHTCOND , PEDROWNOTGRNT , SDOTCOLNUM , SPEEDING , ST_COLCODE , ST_COLDESC , SEGLANEKEY , CROSSWALKKEY , HITPARKEDCAR

Methodology :

In this project python is used for easily avialblity of functionality, coding is performed on IBM watson jupiter notebook. In python data analysis is easy to perform and python also contain sufficient libery for data tranformation like Pandas, Numpy, Matplotlib, and Seaborn .The data was mostly categorical so I stuck to graphical representation to see correlation between various variables.

Process have to be followed for proper prediction severity are :

- Problem Understanding : First understand the problem in user and business aspect. Check if there is already solution of this problem ,if yes then identify the modifications.Otherwise try to find goal and how to reach that goal at least cost.
- Data Collection : Collect the data from standard source or collect data youself with least missing values.
- Data Visualization : Visualise data with matplotlib or seaborn liberay . Try to form correlation with scatter plot , pair plot and heat map ,it will give which feature is important and which have to remove .It also help to identify the outliers.
- Data Transformation : This process involve make the data to satisfies for the mathematical model which have filling missing values, normalised data , try to reduce or remove outlier , remove the independent features.
- Data Modeling : This project target variable is in labeled formed , so supervised learning is used. We have to apply supervised algorithm and get more accuracy.Different algorithm is used in this project random forest , xg boost , svm .
- Data Evaluation : After apply above algorithm we have to check accuracy and have to maximise the accuracy by adjusting paramenter and checking for algorithm. this process done by f1score , confusion matrix , by creating distribution plot ,etc