**CAR ACCIDENT SEVERITY**

SWAPNIL SWARAJ SINHA

AUGUST 28,2020

**INTRODUCTION:**

We will be using the example dataset given by SDOT Traffic Management Division to predict the car accident severity of Seattle. This city is a place where rainfall is experienced for an average of 18 days/month and it also have high level of humidity and it makes it a perfect place for studying car accident severity due to the weather. Seattle had set a goal that by 2030 they'll have 0% of fatal or serious injuries for car accidents. Although these numbers have decreased through the years it seems very unlikely that the city will reach this goal for 2030. So the purpose of this project Y is to predict the severity of this accidents due to several variables so that the city can take actions accordingly.

**Data Description:**

We will be using the example data set given by SDOT Traffic Management Division that contains 194673 rows and 38 columns of data. This columns contains a vast amount of variables such as location, vehicle count, road condition and lightning condition, weather, collision type and many others.The severity of the car accident can be derived from the severity code column.Collision types, date, junction types, whether the collision is due to inattention, whether the driver was under drug or alcohol, weather conditions, road conditions, light conditions, whether or not the pedestrian right of way was not granted, and speeding are the attributes used in the project to determine the severity of car collisions. With these information we'll train the dataset to predict the accident severity due to most of this variables but most of it focused on the weather and road conditions. We'll need to separate this data set to a train dataset and a test dataset to have the best results possible.

***Dataset:***

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

***MetaData:***

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

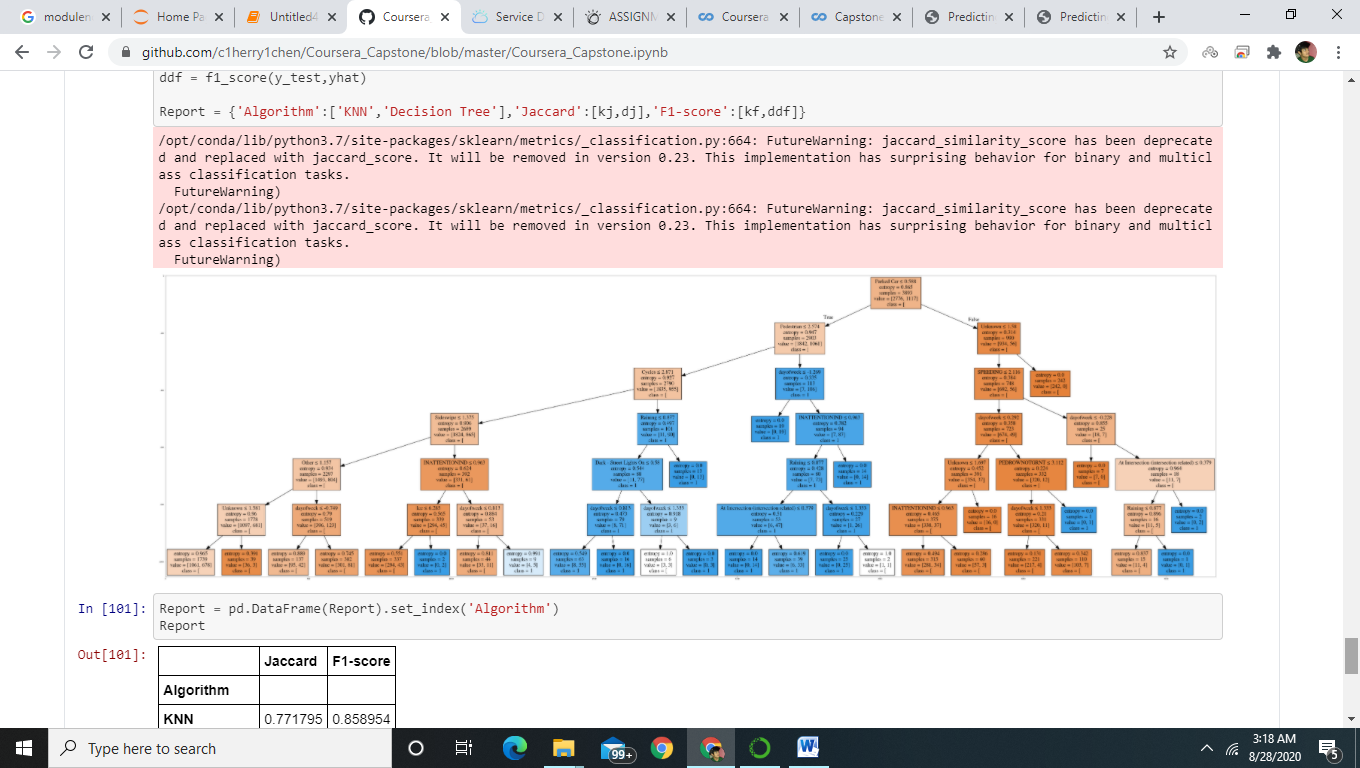
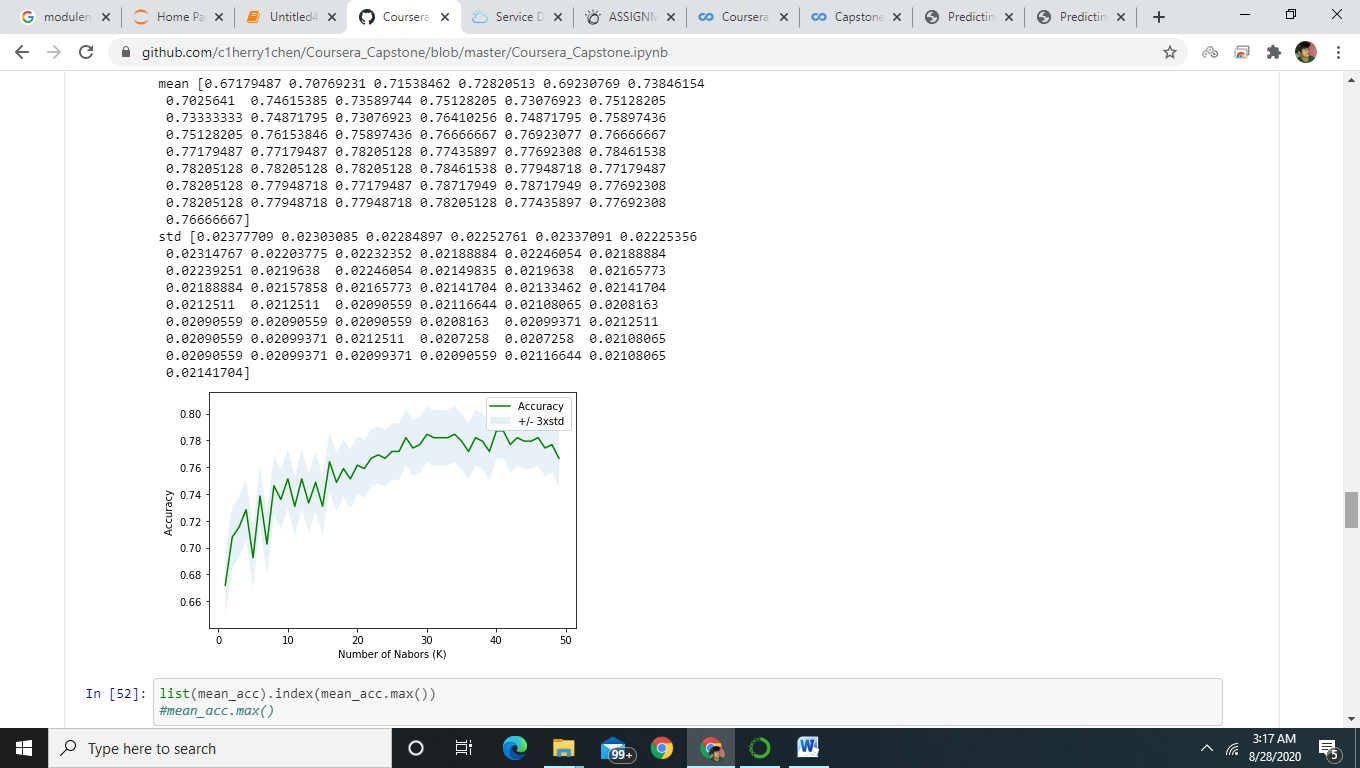
### METHODOLOGY

In the project, the collision data is provided by the coursera capston project as an example.

Firstly, pre-process and clean the data. Irrelevant attributes in the table are dropped, one hot coding is used to convert categorical variables to binary variables. Nan in the columns ['INATTENTIONIND','UNDERINFL','PEDROWNOTGRNT','SPEEDING'] are converted to 0.

Secondly, classification methods including K nearest neighbour (KNN) and decision tree are chosen to build and train the model because there're a lot of attributes in this collision datasets and so, comparing and observe the trend of individual attributes will be tedious and ambiguous. 0.2% and 2% of the data are chosen randomly as test and training data to run the notebook faster.

Finally, jaccard\_similarity\_score and f1\_score are calculated to evaluate the accuracy of the models.Jaccard similarity score is a measurement of the similarity or overlapping of two datasets. It range from 0% to 100% and 100% means two datasets are identical. F1 score is also an accuracy test which compare the true positive and false positive in the predict dataset. When F1-score equals 1, two datasets are identical.



### RESULT

Jaccard similarity score is a measurement of the similarity or overlapping of two datasets. It range from 0% to 100% and 100% means two datasets are identical. F1 score is also an accuracy test which compare the true positive and false positive in the predict dataset. When F1-score equals 1, two datasets are identical.

KNN has Jaccard score 0.772 and F1-score 0.859. Decision tree has Jaccard score 0.779 and F1-score 0.867. Both methods do a decent job in prediction.

### DISCUSSION

Overall, the models built have a good accuracy. Howevever, there are still some improvements we can do in the future. For example, small percentage of the data is chosen to train and predict the model which may affect the accuracy of the results. Although the accuracy of the decision tree is high, it is hard to follow the tree plotted manually. More researches may be required, so that we can drop more features to have a readable tree graph.

### CONCLUSION

The purpose of the project is to build models to predict car collision severity in seattle. K nearest neighbour and decision tree have a high accuracy and are built successfully.