CAR ACCIDENT SEREVRITY PREDICTION

Introduction / Business Problem –

In this project we are attempting to predict the severity of accidents occurring in the US. We will be utilizing the provided data set available in the Kaggle repository using the number of people and vehicles involved in the accidents in the past and what kind of accidents had occurred under which circumstances.

This project can help make decisions to improve the Road Conditions, lighting conditions, etc. which can be considered as couple of factors in the past accidents.

Data Set

The data set used for this project is available at the below link –

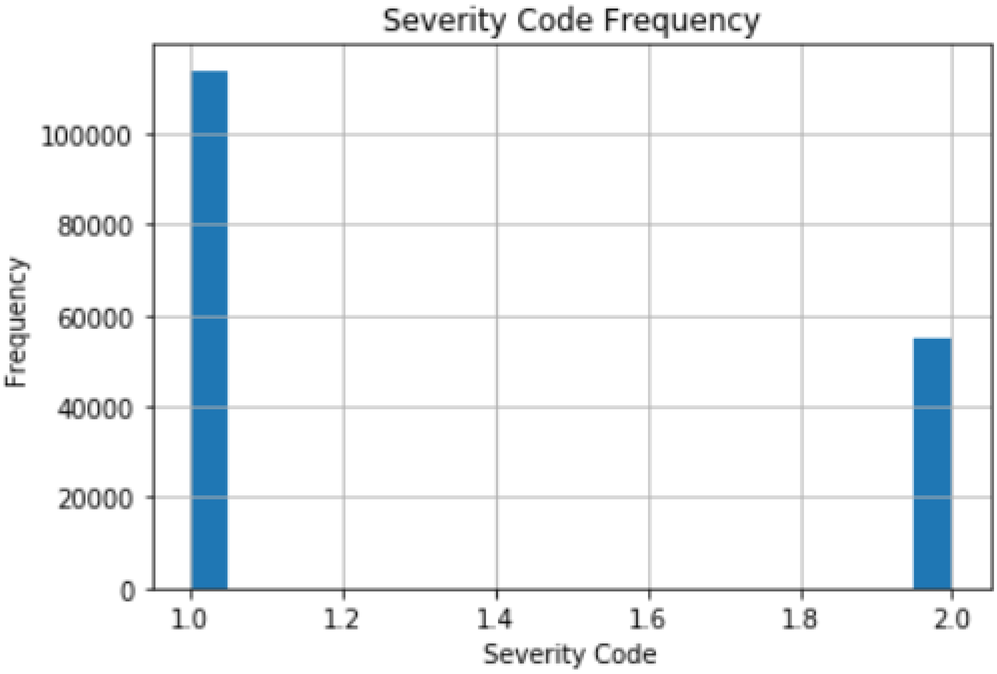
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

This data set contains the driving conditions, the number of people & the vehicles involved n the accidents, and the corresponding severity of the accident.

This data set is an unbalanced data set and many data transformations are required in order to start playing with the data.

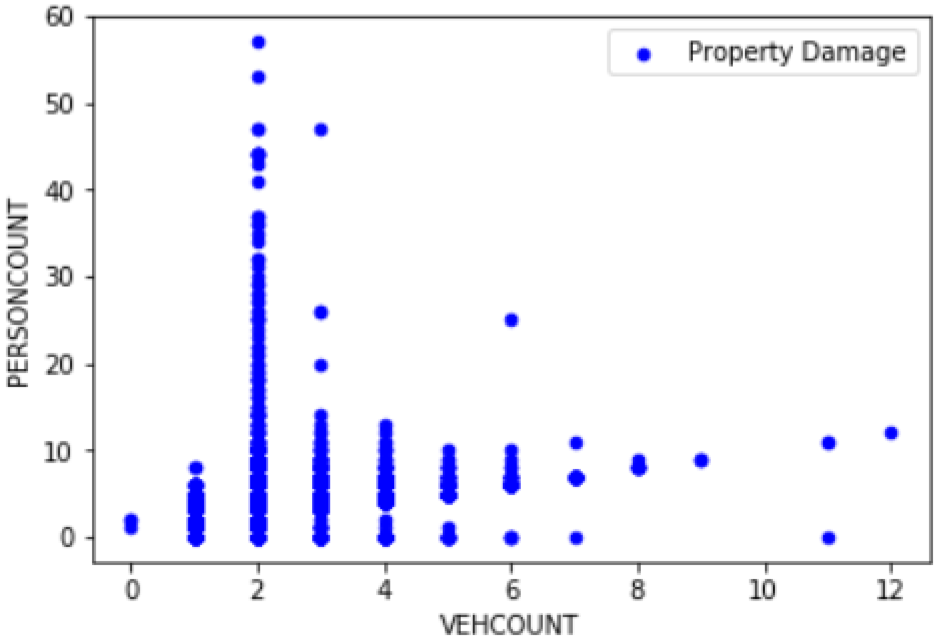
Methodology

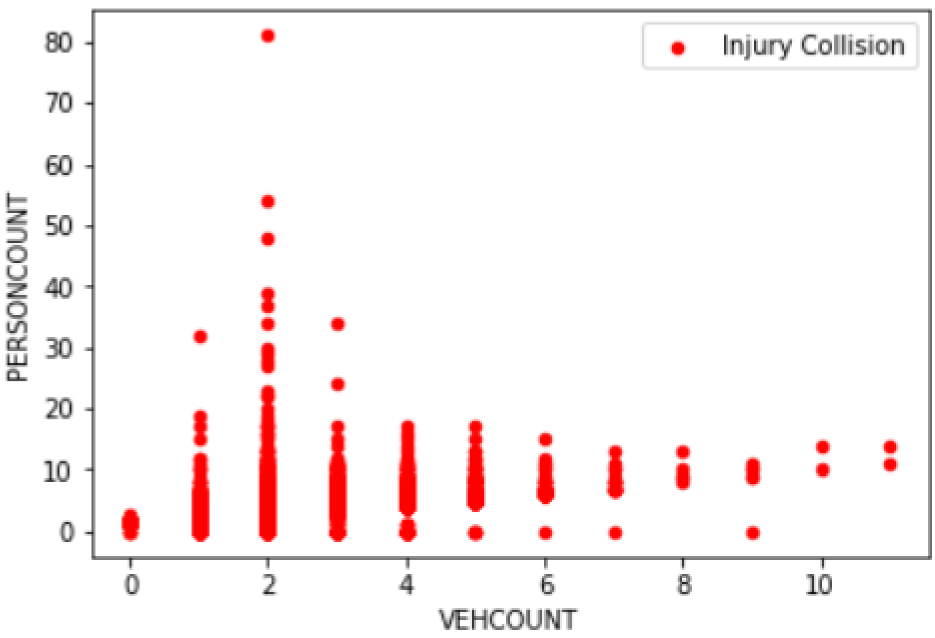
After cleaning the dataset, I plotted three graphs to analyse the data using exploratory analysis. First graph is a simple histogram plotting the severity code against its frequency which resulted in heling us identify that SEVERRITY_CODE = 1 was the most common in the provided data set.



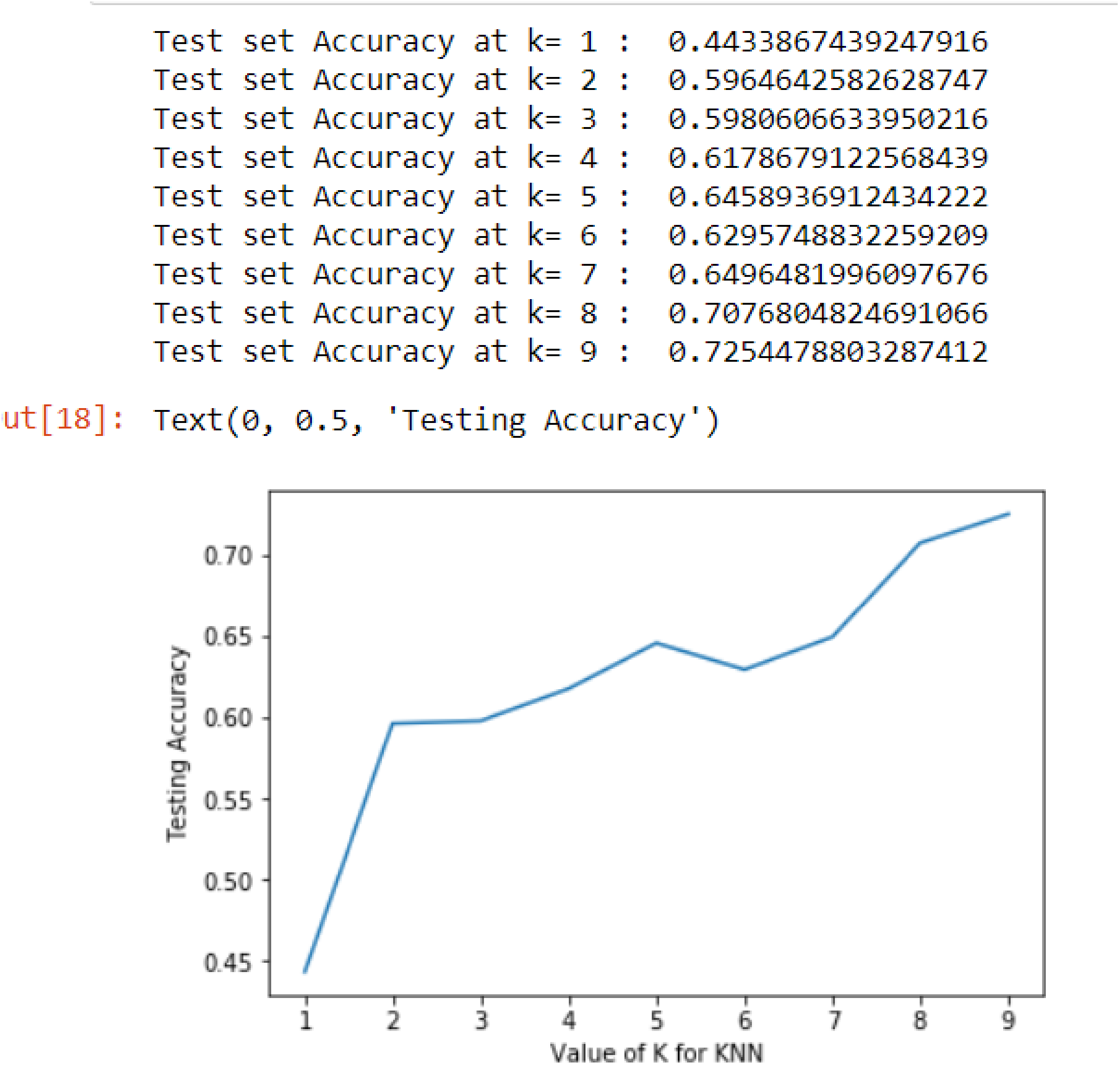
Second graph plotted was to identify the relation between Collision code & frequency. Like the above code, it plots the frequency of Collision Code. In contrast to the severity code, collision code is more specific in stating that the collision code 10 or 11, i.e. either “entering at an angle” or “both going straight, both moving, sideswipe” was the highest amongst the rest.

The last two graphs are scalar graphs representing damage of property and a collision involving human injuries taking into account the number of people and the vehicles involved.





Later, I implemented the KNN Model. First, I tested the accuracy from K=1 to 9 to see which would give me the best results in higher accuracy value.

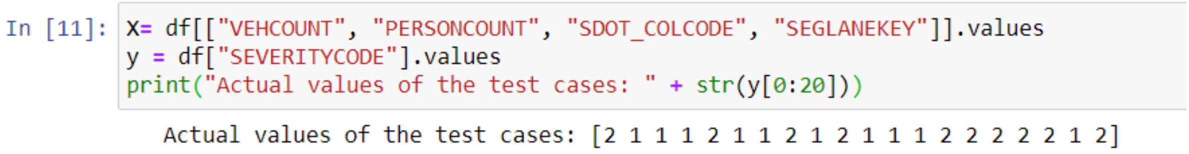


```
Test set Accuracy at k= 1 :  0.44338674392479l6
Test set Accuracy at k= 2 :  0.596464258262874 7
Test set Accuracy at k= 3 :  0.5980606633950216
Test set Accuracy at k= 4 :  0.6178679122568439
Test set Accuracy at k= 5 :  0.6458936912434222
Test set Accuracy at k= 6 :  0.6295748832259209
Test set Accuracy at k= 7 :  0.6496481996097676
Test set Accuracy at k= 8 :  0.7076804824691066
Test set Accuracy at k= 9 :  0.7254478803287412
```

```
ut[18]:  Text(0, 0.5, 'Testing Accuracy')
```

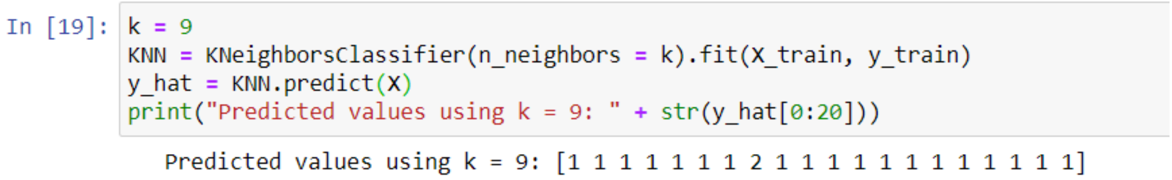
Results

In my analysis, I found that K=9 resulted in highest accuracy value at around 0.72. Then I predicted the value of Y_HAT and it produced 11 correct values out of 20.



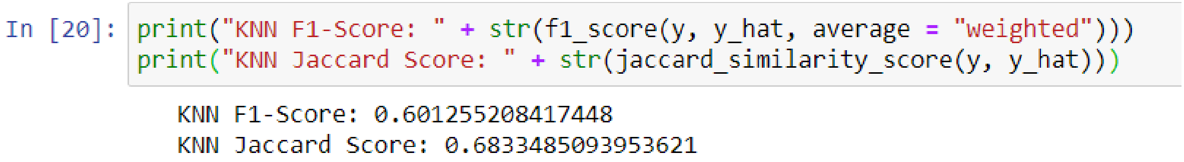
```
In [11]: X= df[["VEHCOUNT", "PERSONCOUNT", "SDOT_COLCODE", "SEGLANEKEY"]].values
         y = df["SEVERITYCODE"].values
         print("Actual values of the test cases: " + str(y[0:20]))

            Actual values of the test cases: [2 1 1 1 2 1 1 2 1 2 1 1 1 2 2 2 2 2 1 2]
```



```
In [19]: k = 9
         KNN = KNeighborsClassifier(n_neighbors = k).fit(X_train, y_train)
         y_hat = KNN.predict(X)
         print("Predicted values using k = 9: " + str(y_hat[0:20]))

            Predicted values using k = 9: [1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1]
```



```
In [20]: print("KNN F1-Score: " + str(f1_score(y, y_hat, average = "weighted")))
         print("KNN Jaccard Score: " + str(jaccard_similarity_score(y, y_hat)))
```

```
KNN F1-Score: 0.601255208417448
KNN Jaccard Score: 0.6833485093953621
```

Using KNN Model, I achieved around 68.3% accuracy.

Discussion

Based on the results above, I believe, if I had incorporated more variables to predict the target variable, the severity, the accuracy would have been higher. Additionally, this project led to thinking that what if instead of predicting the car accident severity after accidented happened, what if we used previous car crashed data using weather, road conditions, lighting conditions, and speeding variables, the likelihood to predict the data that one would get into a car crash given those conditions.

Conclusion

Using all the subjects I learned, I analysed the relationship between the number of people injured, the number of vehicles damaged, kind of collision, and the severity of the collision. I used the KNN model to predict the severity of the CAR accident.