Introduction/Business Section

My idea for a problem is that we can check whether number of persons involved, number of pedestrians involved, number of cycles involved and number of vehicles involved influence the severity type of an accident. This data science problem is directed towards the state police and traffic police who would like to know which of the above is affecting the severity of an accident and hence be cautious about it. They might regulate the traffic of vehicles on basis of this data or even advise their citizens of some new policies.

Data section

I will be using the accident data provided by Coursera.

It has a column indicating severity type and 4 columns indicating number of persons, pedestrians, cycles, vehicles involved.

Methodology

I would like to predict the severity type of an accident based on the number of persons, pedestrians, cycles or vehicles involved. This is a classification problem. So,

1. I chose to perform k-nearest neighbour algorithm on the available data.
2. I stored the 5 columns accordingly into a new data frame in a python file.
3. Then standardized the independent variable values using StandardScaler.
4. Then I split the data into training and test data using train_test_split.
5. Then using python, I checked the accuracy scores for different values of k (number of nearest neighbours) from 0 to 11.
6. Then taking the optimal value of k, I performed the algorithm again and tested it on my test data.

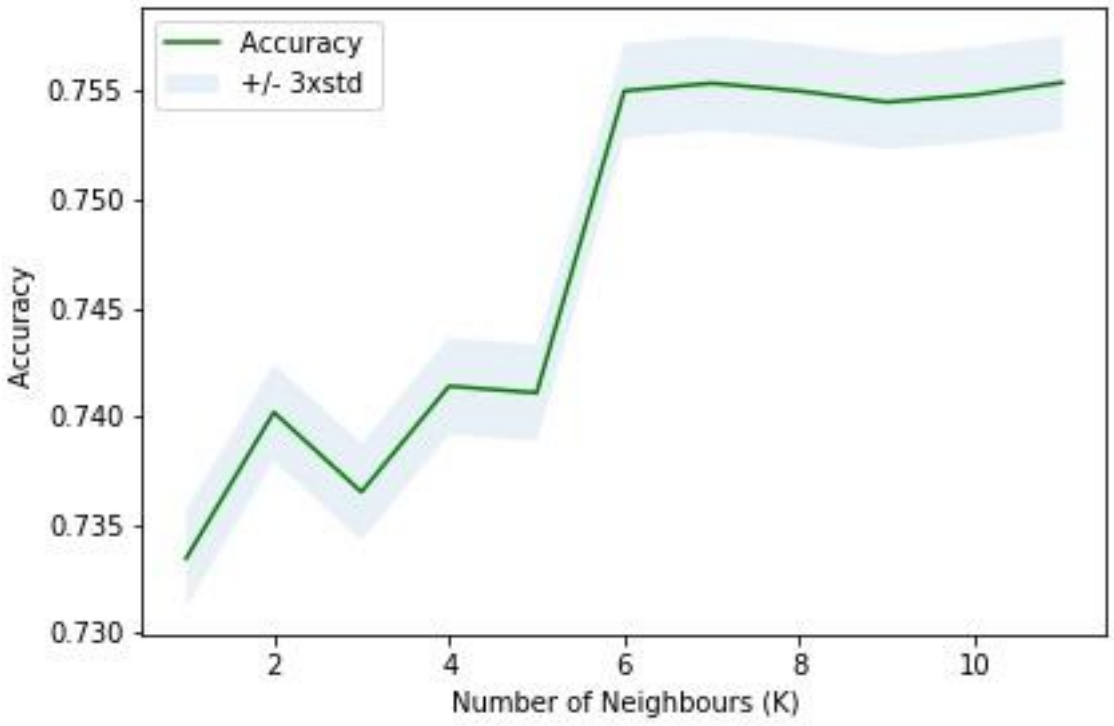
Results

After running the k-nearest neighbour algorithm for 12 values of k , we get the following values of accuracy scores:

```
0.73345319, 0.74018236, 0.73650957, 0.7413895 , 0.74108129, 0.75500193,
0.7553615 , 0.75500193, 0.75448825, 0.75482214, 0.75538718
```

We see that the increase in accuracy score is very subtle after k=7. It even drops after k=7. So, we choose k=7 as optimum.

Then we get the plot for different k values.



Then we train the model with k=7 for the training data and predict severity type for both training and test data. First 19 values are shown for the test data and its predicted counterpart.

```
[1 1 2 2 1 2 2 2 1 1 2 1 1 1 1 1 1 2 1]
[1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1 1]
```

Then we check the accuracy scores of them both:

```
Train set Accuracy:  0.7522313115617255
Test set Accuracy:  0.7553614999357904
```

We note that the accuracy of predicting is even better for the test data!

So, our model is working significantly well.

Discussion

As per the results, we can predict the severity type of an accident based on the number of persons, pedestrians, cycles or vehicles involved as the accuracy scores are quite high.

Conclusion

K-Nearest Neighbour is a good method of predicting severity type in this case, as we see from the results.