Coursera Capstone: Accident Severity (Seattle)

Introduction

Background

Each year, 1.35 million people are killed on roadways around the world. Every day, almost 3,700 people are killed globally in road traffic crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians. More than half of those killed are pedestrians, motorcyclists, and cyclists. In USA in 2019, an estimated 38,800 people lost their lives to car crashes and about 4.4 million people were injured seriously enough to require medical attention in crashes.

Our focus in this project will be on the city of Seattle, Washington.

Business Problem

Any government would try to reduce the number accidents happening under their jurisdiction as road accidents pose a public health and development challenge and greatly effect the human capital development

The data was collected by Seattle Spot Traffic Management System and is available on the Seattle Geodata website. The data we have is dated from October, 2003 to September,2020. It contains information such as severity code, location, type of collision, number of people involved, etc.

The target audience for this report is the Local Government of Seattle, the first responders, insurance companies and the general public as well. The result will provide the target audience insightful data to reduce the number of accidents and be able to handle accidents more efficiently in Seattle.

Data Acquisition and Cleaning

Data Sources

The entire dataset was downloaded from the Seattle Geodata website and can be downloaded from here.

Data Cleaning

The raw data contained 221,525 rows and 40 columns but the data contained a lot of problems

First, a lot of columns contained unique values (eg. INCKEY, OBJECTID, etc) and each of these had to be identified individually and removed as these columns do not help with any kind of visualisation and don't contribute with the prediction algorithm.

Second, there were columns (INATTENTIONID, PEDROWNOTGRNT, SPEEDING, etc) which have an excessive amount of missing values and cannot we use at all

Third, columns with repeating values (INCDATE, SEVERITYDESC and SDOT_COLCODE) were removed.

Fourth, columns for latitude and longitude (X and Y) were used for making heatmaps but were removed for the prediction algorithm development.

Fifth, the columns INJURIES, SERIOUSINJURIES and FATALITIES were dropped because this is what we have to predict using our algorithm and having them as input for our prediction algorithm will defeat the purpose.

Sixth, separate data frames were made using columns from the main data frame for making maps and graphs while the main data frame was used for prediction algorithm development but in all the cases all rows with missing values were just removed as ample of data was still left after removing these rows and using predicted values would have tempered with the results.

Finally, we were left with 188,255 rows and 17 independent variables/ columns for the algorithm development.

Exploratory Data Analysis

Mapping

Since we are talking about accidents there are bound to be hot zones where more number of collisions happen, so the following maps will be show casing the areas where more precautions need to be taken and extra measures need to be taken to decrease collisions.

- First, I started with a heatmap for all type of collisions and the result was a red layer covering the entire city of Seattle so I decided to see the hot zones for every year from 2004 to 2019 and I saw that downtown Seattle has had the most number of road accidents ever year without fail.

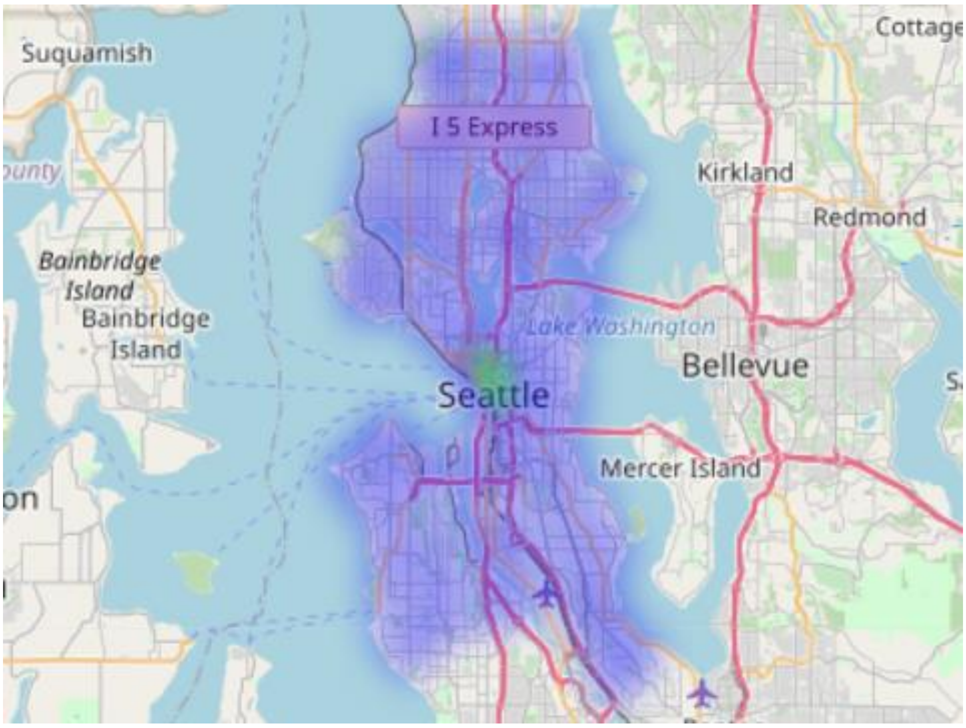


Figure-1

- Second, I made a heatmap to show where cycle related accidents are more likely to happen as the city of Seattle has poured millions into making it more bike friendly and was ranked seventh safest city for cyclists in USA by CNBC in December of 2019.

The map shown in figure-2 is similar to the heatmaps generated during year by year analysis which shows that more work needs to be done in the following areas as they have been the hot zone for accidents for years.

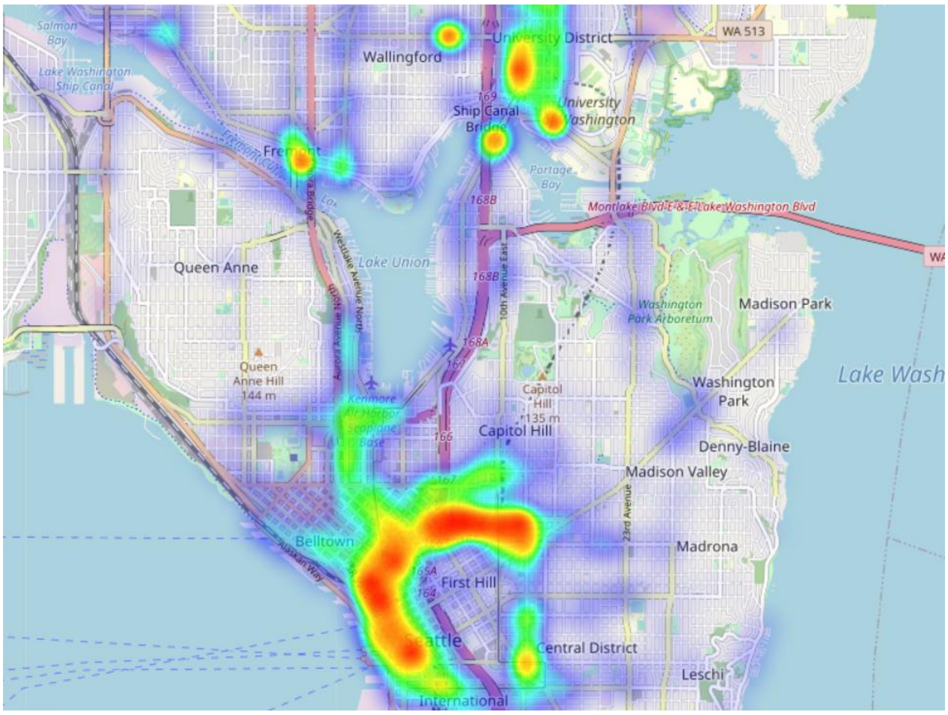


Figure-2

- Third, I went for a heatmap to see if accidents where intoxication was a factor can be pin pointed to particular locations and downtown was highlighted again but upon more research, I found out that there is a high density of bars in highlighted areas

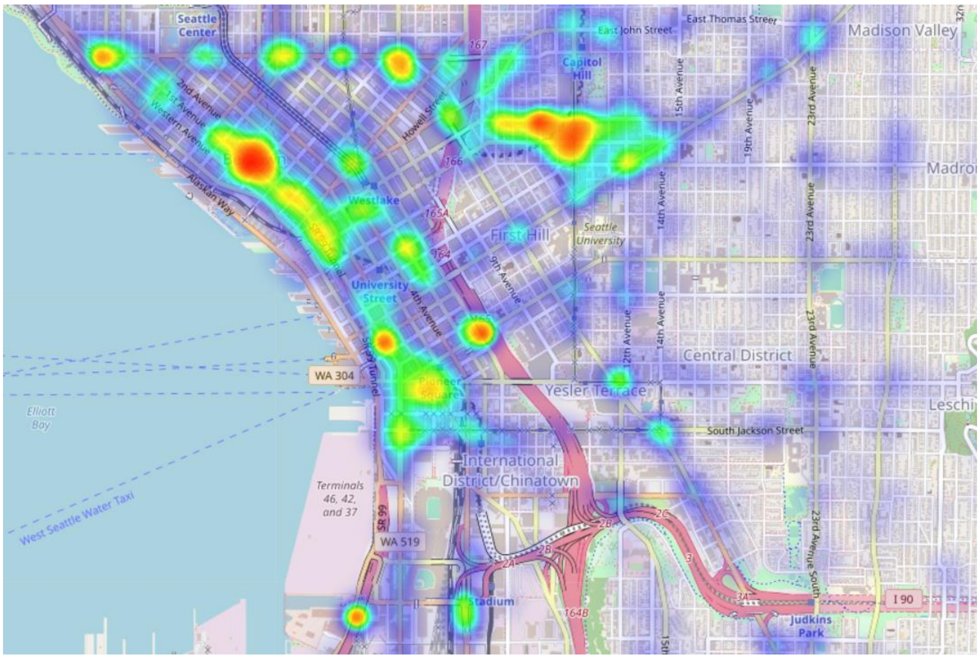


Figure-3

Graphical Representation

- First, I start with a graph showing the overall trend of number of accidents every year since 2004 to 2019 and as shown in the graph below, in the last 5 years number of accidents have seen a decline from 14,260 in 2015 to 11,204 in 2019. Although number of accidents have decreased over the years, Seattle is still one of the worst cities to drive in USA according to multiple independent studies.

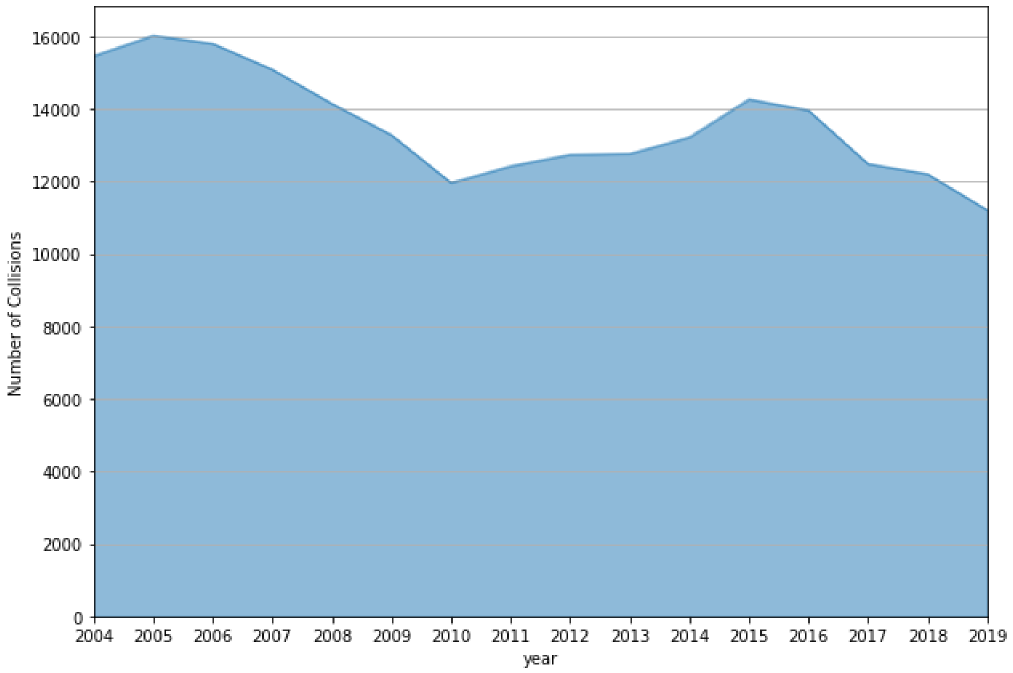


Figure-4

- Second, I tried to find some kind of pattern in number of fatalities over the years but nothing can be interpreted from the following graph except that 2019 saw a sharp rise in number of fatalities with 26 being reported in 2019 as compared to 14 being reported in 2018. In 2019, seattle also saw the maximum number of road accident related fatalities in over 10 years.

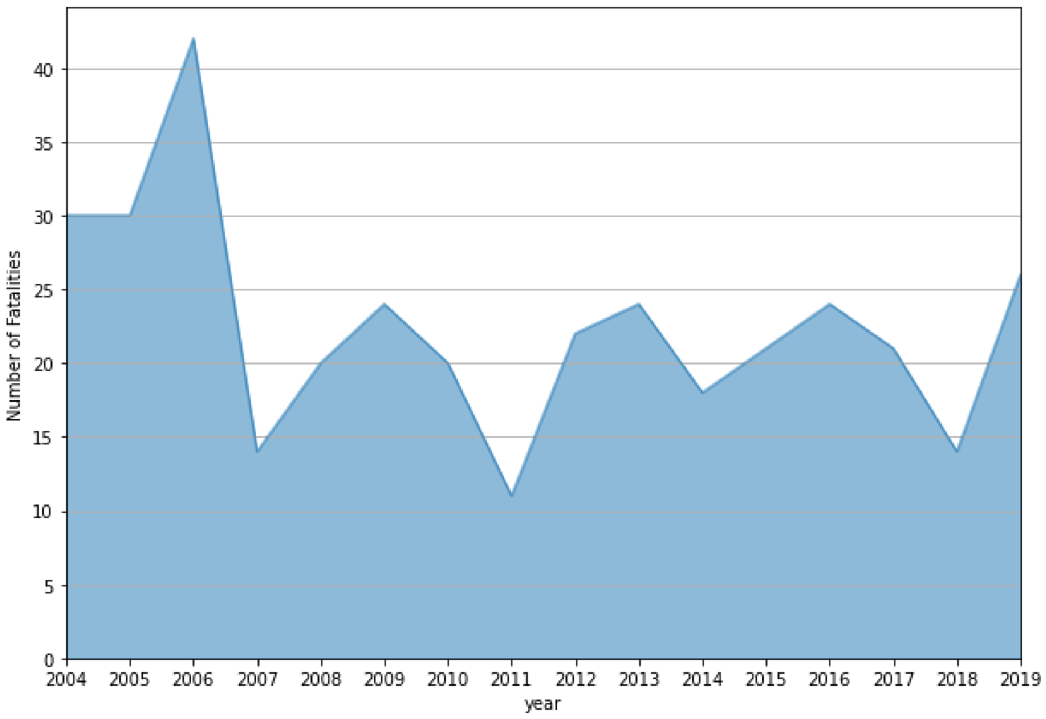


Figure-5

- Third, according to NHTSA (National Highway Traffic Administration), most accidents tend to occur between 3 pm to 6 pm. This is often because of the high volume of vehicles on the road. The following graph for the city of Seattle shows a similar trend with accidents peaking around 5pm.

  Also, as visible in the figure-6 a trend similar to the overall trend is shown by the data of 2019.

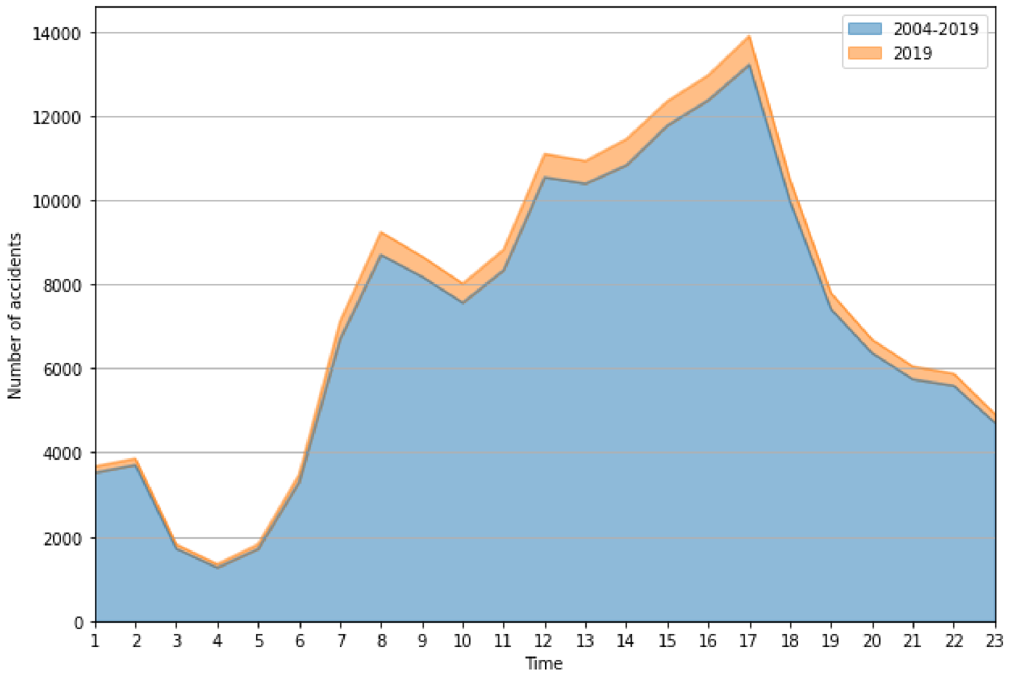


Figure-6

- Fourth, according to NHTSA Saturday is considered to be the most dangerous day to drive in USA but according to the graph shown below people are more likely to get into a road accident on weekdays with Friday and not Saturday being the most dangerous day.

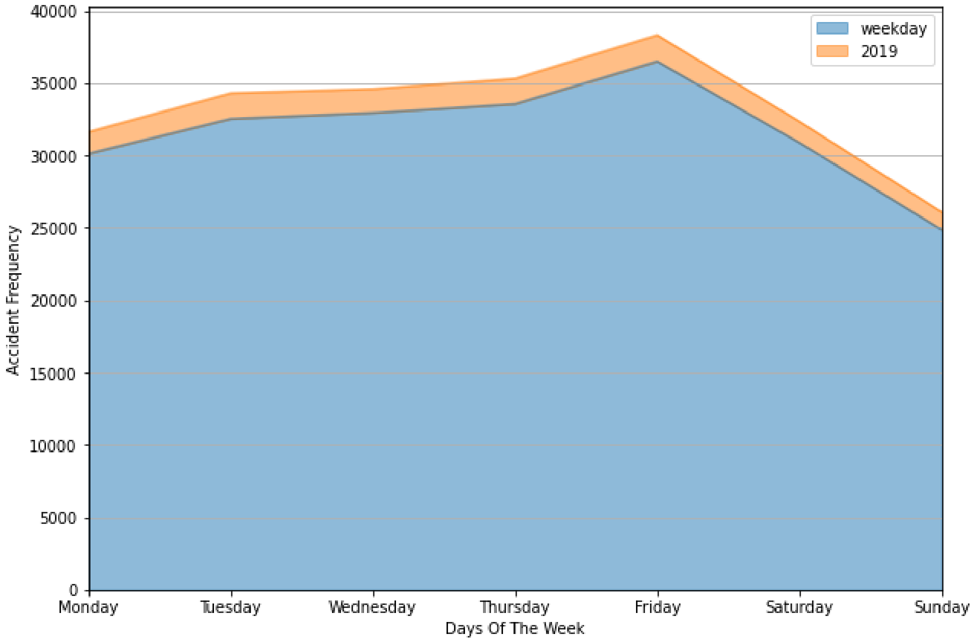


Figure-7

- Fifth, collisions are also of different types and to reduce the number of collisions we need to identify and tackle each of them separately. In the data, 9 main types of collisions were identified and other minor types of collision were simply labelled as others.

  According to NHTSA, in most type of collision is rear ended but as shown in the pie chart, in Seattle, most collisions involve parked cars.

  Also, another observation to be made is that people are more than four times likely to have a collision while taking a left turn as compared to while taking a right turn.

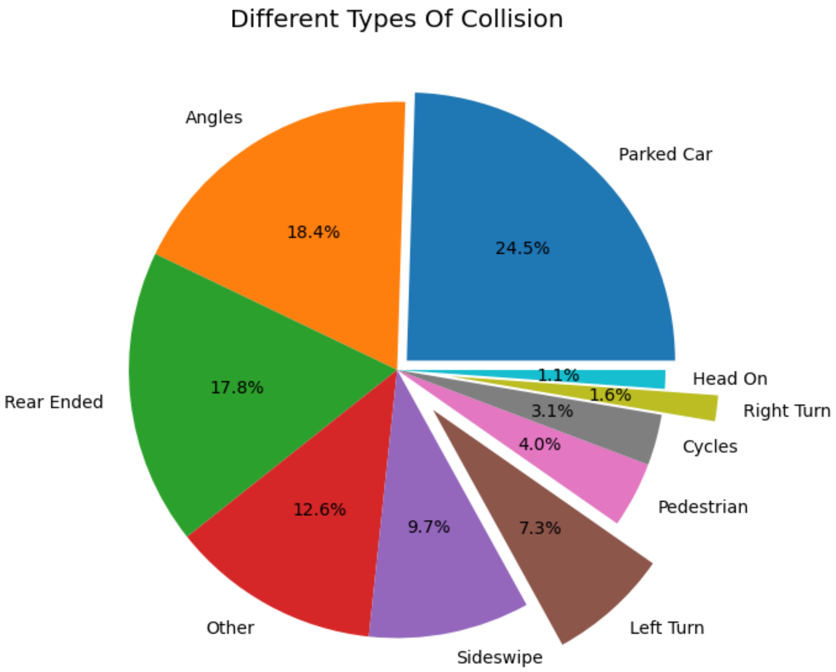


Figure-8

- Sixth, often cyclist follow the same precautions that are given for motorized vehicles to follow but although both of them are mode of transportation, cyclists need to take some different precautions to take safely and these things also need to be acknowledged by the government.

  As visible in the following pie charts cyclists are more likely to get into an accident in an intersection while motorists are more likely to get into an accident on a block

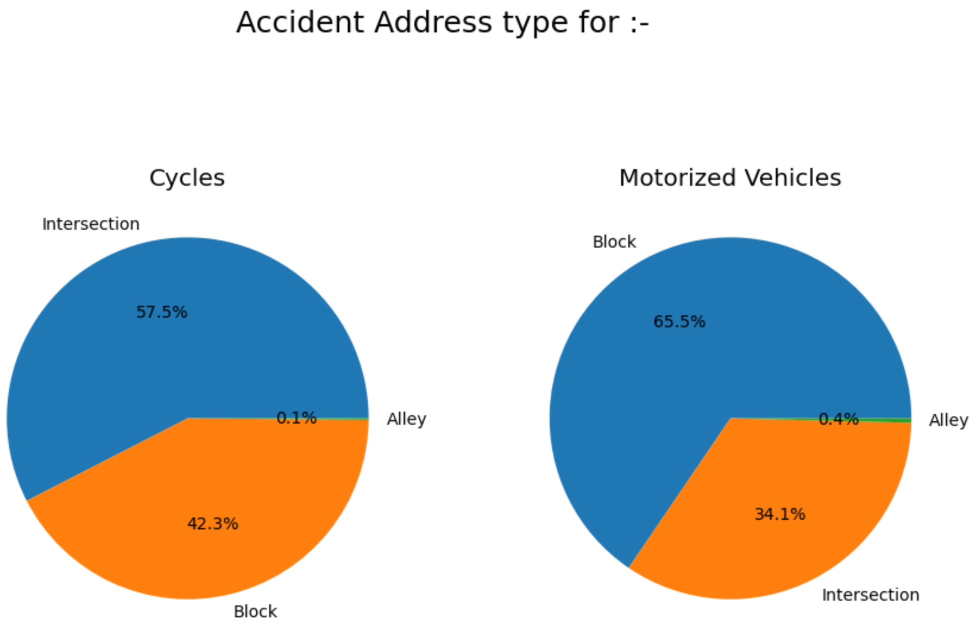


Figure-9

Predictive Modelling

Overview

This is a multi-classification problem, where based on the given variables, the severity of a collision will be predicted. So, the machine learning algorithms used are: -

1. Decision Tree Classifier
2. K Nearest Neighbors
3. Naïve Bayes

We used an 80-20% train-test-split of data. The test and train dataset were almost balanced as shown below.

Train Dataset	
Property Damage	67.76%
Injury	30.45%
Serious Injury	1.60%
Fatality	0.18%

Test Dataset	
Property Damage	67.76%
Injury	30.45%
Serious Injury	1.61%
Fatality	0.17%

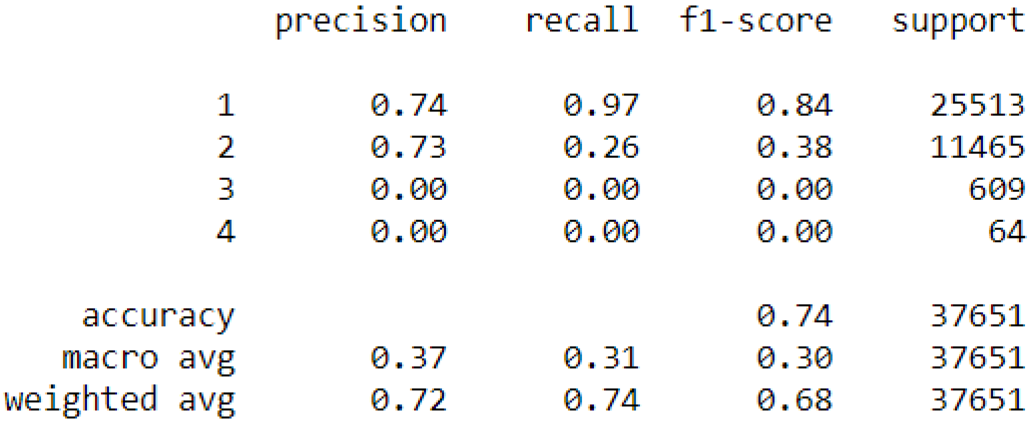
Table 1 & 2

Decision Tree Classifier

A decision tree is a machine learning algorithm that partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. Various branches of variable length are formed.

The goal of a decision tree is to encapsulate the training data in the smallest possible tree. The rationale for minimizing the tree size is the logical rule that the simplest possible explanation for a set of phenomena is preferred over other explanations.

Classification report: -



```
              precision    recall  f1-score   support

           1       0.74      0.97      0.84     25513
           2       0.73      0.26      0.38     11465
           3       0.00      0.00      0.00       609
           4       0.00      0.00      0.00        64

    accuracy                           0.74     37651
   macro avg       0.37      0.31      0.30     37651
weighted avg       0.72      0.74      0.68     37651
```

K Nearest Neighbors

K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

Validation Error Curve: -

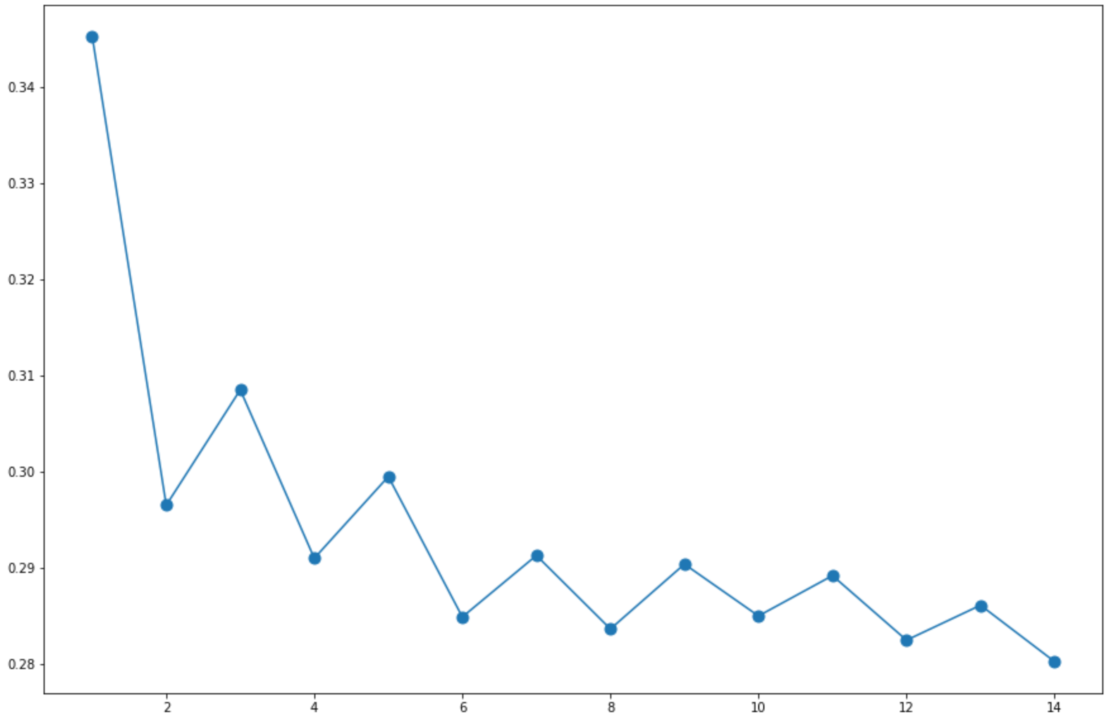
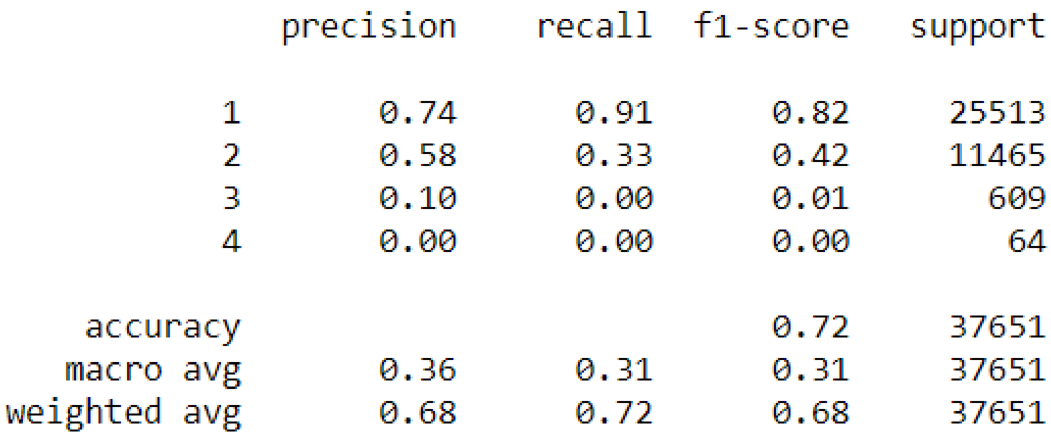


Figure-10

The value of n_neighbors is determined using this curve and value at the elbow(in this case 6) is considered to be the optimum value

Classification Report: -

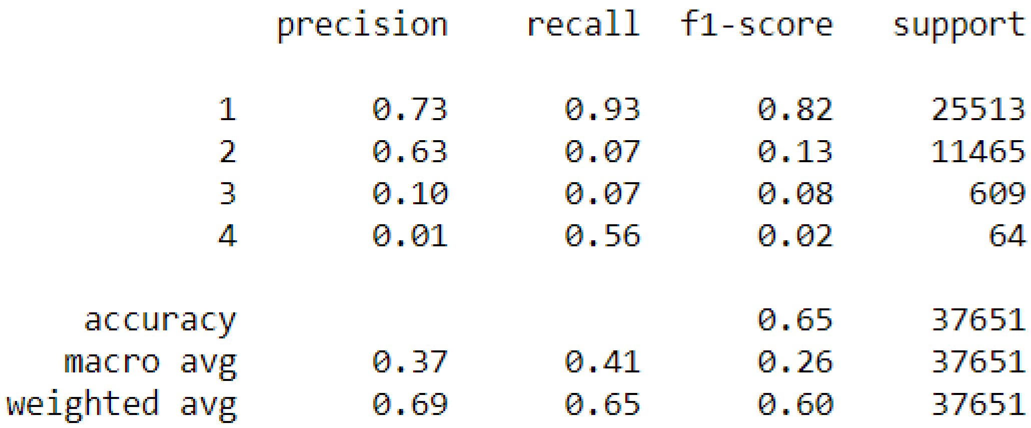


```
              precision    recall  f1-score   support

           1       0.74      0.91      0.82     25513
           2       0.58      0.33      0.42     11465
           3       0.10      0.00      0.01       609
           4       0.00      0.00      0.00        64

    accuracy                           0.72     37651
   macro avg       0.36      0.31      0.31     37651
weighted avg       0.68      0.72      0.68     37651
```

Naïve Bayes

Naive Bayes is a collection of classification algorithms based on Bayes Theorem. It is not a single algorithm but a family of algorithms that all share a common principle, that every feature being classified is independent of the value of any other feature.

Classification Report: -



```
              precision    recall  f1-score   support

           1       0.73      0.93      0.82     25513
           2       0.63      0.07      0.13     11465
           3       0.10      0.07      0.08       609
           4       0.01      0.56      0.02        64

    accuracy                           0.65     37651
   macro avg       0.37      0.41      0.26     37651
weighted avg       0.69      0.65      0.60     37651
```

Performance Comparison

The accuracy of the built models using different evaluation metrics

Algorithm	F1 Score(weighted)	Jaccard Score
Decision Tree Classifier	0.682782	0.735279
K Nearest Neighbors	0.681620	0.715094
Naïve Bayes	0.596810	0.654910

Conclusion

In this project, I analysed the relationship between accident severity and factors like junction type, vehicle and pedestrian count, time, etc. I made heatmaps and graphs which show data which can be used by the authorities and citizens of Seattle and I made classification models which can be used by first responders to assess a situation in case of less information being conveyed to them and react accordingly.