# Predicting the occurrence of severe car accidents in Seattle, Washington, USA

## 1. Introduction

### 1.1 Background

Road traffic accidents and subsequent injuries have a remarkable social impact for individuals and families and are one of the leading causes of death across the globe. Although Seattle in one of the safest cities in the United States of America (USA), according to Seattle government more than 10000 collisions a year still occur, resulting in an average of 20 people losing their lives and 150 people being seriously injured. During the first half of the year 2019, 101 people were seriously injured or killed in 98 collisions on Seattle streets. That is the highest number of crashes in the first six months of a year since 2010, according to preliminary police reports analysed by the Seattle Department of Transportation (SDOT). The city of Seattle is aiming at ending traffic deaths and serious injuries on city streets by 2030, however, some reports argue that if better measures don't take place Seattle will be far from meeting its goal.

Seattle has a warm temperate oceanic climate characterized by cool, wet winters and mild, relatively dry summers. Summertime is mildly warm and moderately dry with temperatures ranging between 20-23 Celsius degrees, with daylight hours up to sixteen hours. In contrast, winters are cold and wet, with low temperatures between 2-10 Celsius degrees, and due to the rainfall most days roads are wet and slippery. Rainfall is experienced for an average of eighteen days per month, and the rain gauge collects approximately 147.3 mm. Snow falls are high and regular, accumulating up to 43.2 mm, mostly in December and February. The average sunshine hours per day are two. Adverse conditions and road status could play a pivotal road in the occurrence of fatal car accidents.

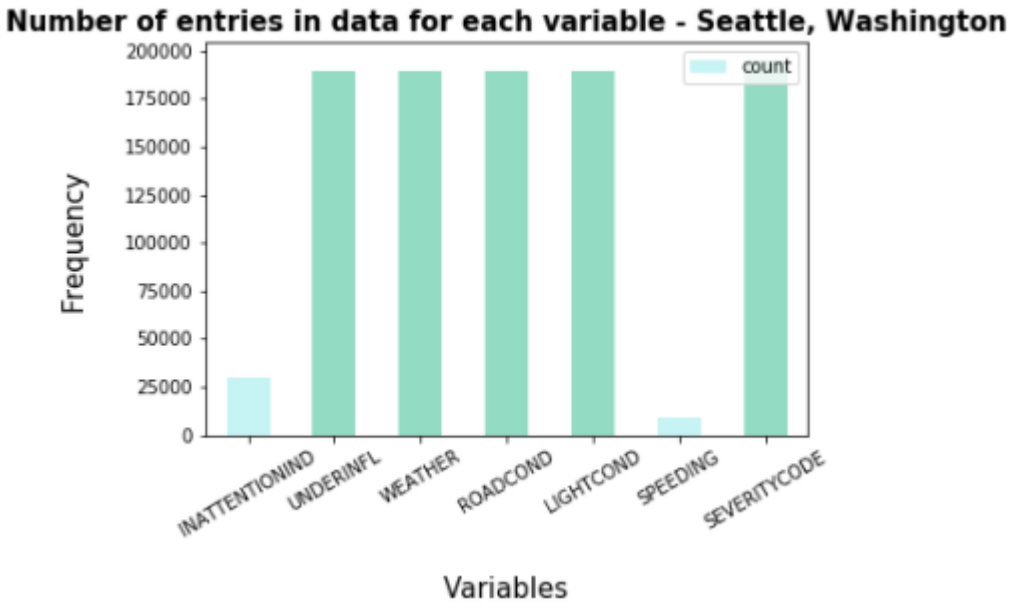
### 1.2 Problem and Interest

In order to establish new measures to reduce the number of mortalities and injuries caused by car collisions, this study aims to find features like road conditions, light conditions, speed, weather, under influence of drugs and/or alcohol, and inattention that can foster car collisions, with the ultimate goal of preventing or diminishing car crashes, making Seattle's roads safer for all inhabitants, extendable to other cities across USA and the rest of the world.

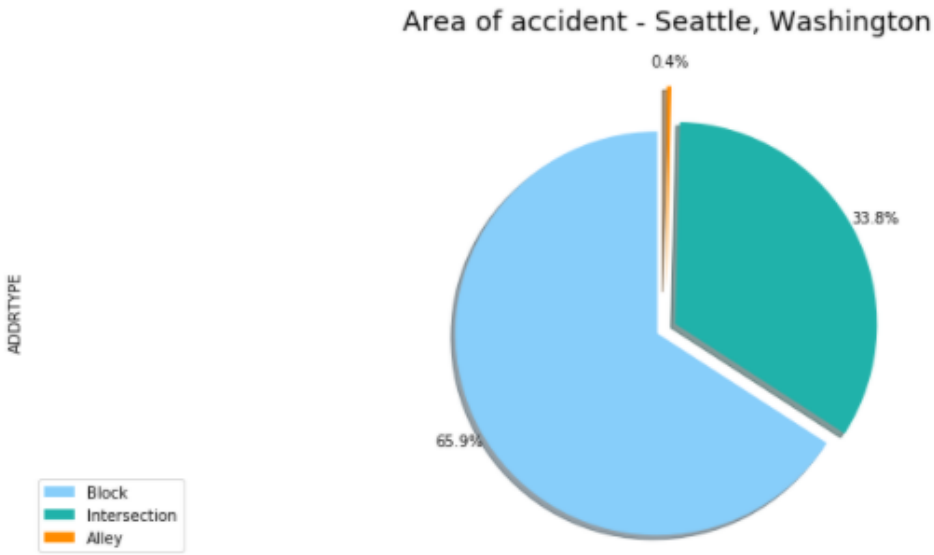
## 2. Data

The data used for this study comprises a csv file and metadata from all collisions in Seattle provided by the Seattle Police Department (SPD) and recorded by Traffic Records from 2004 to present. I will use the data to structure a classification model and determine whether the above mentioned features can promote car collisions leading to injured people.

The models reported herein aim to predict the severity of an occurred accident, represented by the variable Severity Code (1 = Property Damage Only) and 2 (Physical Injury). Such groups were encoded to the form of 0 (Property Damage Only) and 1 (Physical Injury). Following that, 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either "Other" or "Unknown", deleting those rows entirely would have led to a lot of loss of data which is not preferred.




In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had “Other” and “Unknown” in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.




As seen in the plot, most of accidents occurred around blocks.

Features selected were as follows:

- INATTENTIONIND: Whether or not the drive was inattentive (Y/N)
- UNDERINFL: Whether or not the driver was under the influence (Y/N)
- WEATHER: Weather condition during the collision (Overcast/Rain/Clear)
- ROADCOND: Road condition during the collision (Wet/Dry…)
- LIGHTCOND: Light conditions during the collision (Lights on/Dark with light on)
- SPEEDING: Whether the car was above the speed limit at the time of collision (Y/N)

## 3. Methodolody

The Machine Learning models used in this study were:

- **Decision Tree Analysis**: The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- **Logistic Regression**: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable
- **K-Nearest Neighbor (KNN)**: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance)
- **Support Vector Machine (SVM):** The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N — the number of features) that distinctly classifies the data points.

## 4. Results

Results from each of the four models were:

| Althorithm | Accuracy | f-1 score | Jaccard Index |
|---|---|---|---|
| Decision Tree | 0.696 | 0.57 | 0.7 |
| Logistic Regression | 0.696 | 0.57 | 0.7 |
| KNN | 0.694 | 0.58 | 0.69 |
| SVM | 0.696 | 0.57 | 0.7 |

Decision Tree: The criterion chosen for the classifier was entropy and the max depth was 6

Logistic Regression: The C used for regularization strength was C = 0.01 and solve = liblinear

k-Nearesy Neighbor: The best K for the model was k=4

SVM: The C used was C = 1.0

## 5. Conclusion

When comparing all the model's accuracy, we can see that they are pretty similar. Likewise, f-1 score and Jaccard Index show similar values. Logistic regression, Decision Tree and SVM have the exact value of accuracy, f-1 score and Jaccard Index. It can be concluded that the three models can be used to obtain the best performance.

## 6. Recommendations

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.