A Safer Seattle

Using Machine Learning to Identify Risks for Severe Accidents and Inform Decisions to Mitigate Those Risks

---

Douglas Bryan Daniels, Jr.

September 23, 2020

1. Introduction

In 2004, the mayor of Seattle prioritized the minimizing of severe traffic accidents in the city and as a result, Seattle has seen a decline in such accidents over the past few years. However, the mayor wants to see further improvements and is calling for a new study that will guide decision making in the coming years. His urgency is now peaked as the Seattle Area Department of Economic Development recently proposed having a week-long festival in 2022 called ***SuperSonic***. The aim of the festival is to attract investors, corporations, and international entrepreneurs to Seattle in hopes that they will invest and even relocate to Seattle.

Realizing that a safer driving record will enable Seattle to better promote itself to achieve its goals for ***SuperSonic***, the mayor has assembled a task force to determine the variables most closely linked to severe accidents and to propose ways to mitigate the risks. While the most pressing purpose is to take steps to reduce severe car accidents before 2022, the task force has been asked to also consider any proposals that would take longer to implement to achieve improved transportation safety over the long term.

Upon receiving the report, the mayor will present the findings to the appropriate city officials for evaluation and implementation.

2. Data Acquisition and Cleaning

The data set used by the task force for this project was provided by the Seattle Police Department and was recorded by Traffic Records. The data set includes all types of collisions from 2004 to present, and it can be accessed here.

The data was obtained in CSV format, and it contains 37 features, along with a severity rating for each of the 194,673 accidents as follows:

Code	Description
3	fatality
2b	serious injury
2	injury
1	property damage
0	unknown

***It should be noted that the data that we received for the study did not match the specifications provided, as there were no accidents included with a severity rating of 3 (fatality) or 2b (serious injury). Under normal circumstances, the task force would have contacted the Seattle Police Department to determine why this information was excluded from the dataset we received.**

The data features include location of the accident, date, weather conditions, road conditions, whether or not the vehicle involved was speeding, etc.

While there may have been other accidents in the Seattle area were not recorded by Traffic Records, this report is official and we can reasonably assume that this is the most exhaustive, accurate source for the information we are seeking. The cost/benefit ratio of obtaining relevant, supplemental data sources would be far too high to pursue.

Since our plan is to reduce the severity of traffic accidents and our accidents are labeled with accident severity classifications, we will employ a supervised classification learning model. With that in mind, let us consider what actions we need to take with regard to understanding and preparing our data set for modeling.

We will begin by observing the data set and determining which attributes we will use to train our model. This will include removing features that are deemed irrelevant in predicting the severity of an accident. We will also remove accident records with too much missing data or utilize appropriate statistical methods to fill in missing data where appropriate. Duplicate, highly similar and highly correlated features will be removed, along with features that contained mostly null values with no way of interpreting their actual values.

A small number of accidents were missing location values. Because these rows often had a significant number of other missing data and represented such a small proportion of the whole dataset, they were removed.

Because weather conditions are of specific interest in this study, we will limit the cleaned dataset for our models to only 5 features: X-coordinate, Y-coordinate, weather conditions, speeding, and under the influence. We will also limit our dataset to the last 3 complete years of data: 2017-2019.

Given that the most severe traffic accident cases are generally a very small proportion of total traffic accidents, we will need to balance the labeled data so as not to bias the model.

At this point we will move on to exploratory data analysis, modeling and evaluation, looping back as needed to perform further work on the data.

3. Exploratory Data Analysis

Folium maps were used for the visualization of spatial data. A mark cluster object was added tot he map to superimpose locations of 2017-2019 accidents on the map of Seattle. City-wide it was observed that a very high proportion of accidents along the I-5 corridor running North-South through the city (Figure 1).

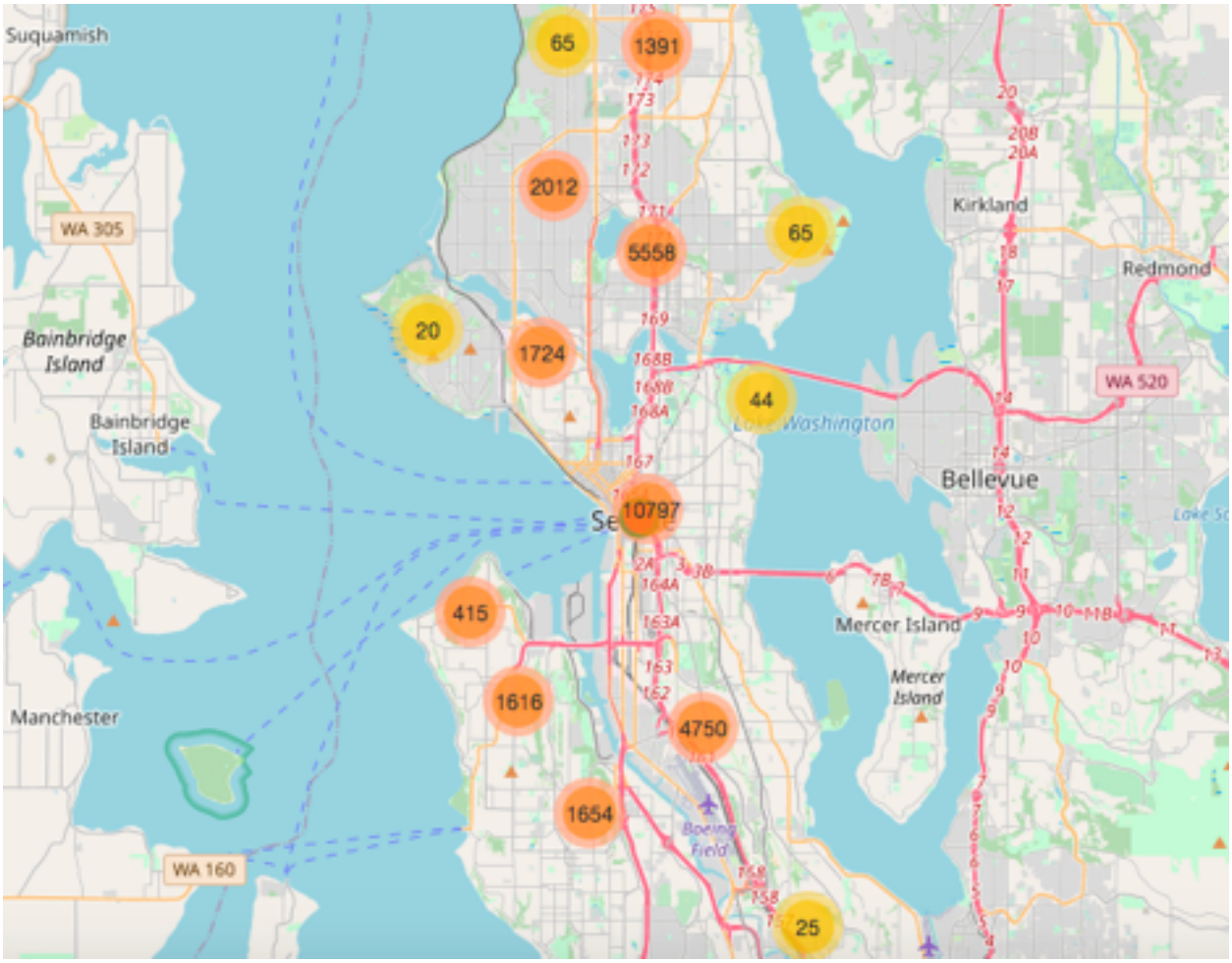

Figure 1. 2017-2019 Traffic Accident Locations

Data on traffic accident numbers at intersections throughout the city revealed that the James St.-6th Avenue intersection in 2017-2019 saw nearly 65% more accidents than any other single intersection in the city. The weight of this finding is increased by the fact that no other intersection in the city exceeded trailers in accident numbers by more than 7%.
In addition, 45% of the accidents at this intersection were severe, compared to 32% of all accidents city-wide classified as severe.
Exploration of data extending back to 2004 reveals that this trend is not unique to the 2017-2019. This intersection has been highlighted in Figure 2.

Figure 2. James St. - 6th Ave. Intersection in Downtown Seattle

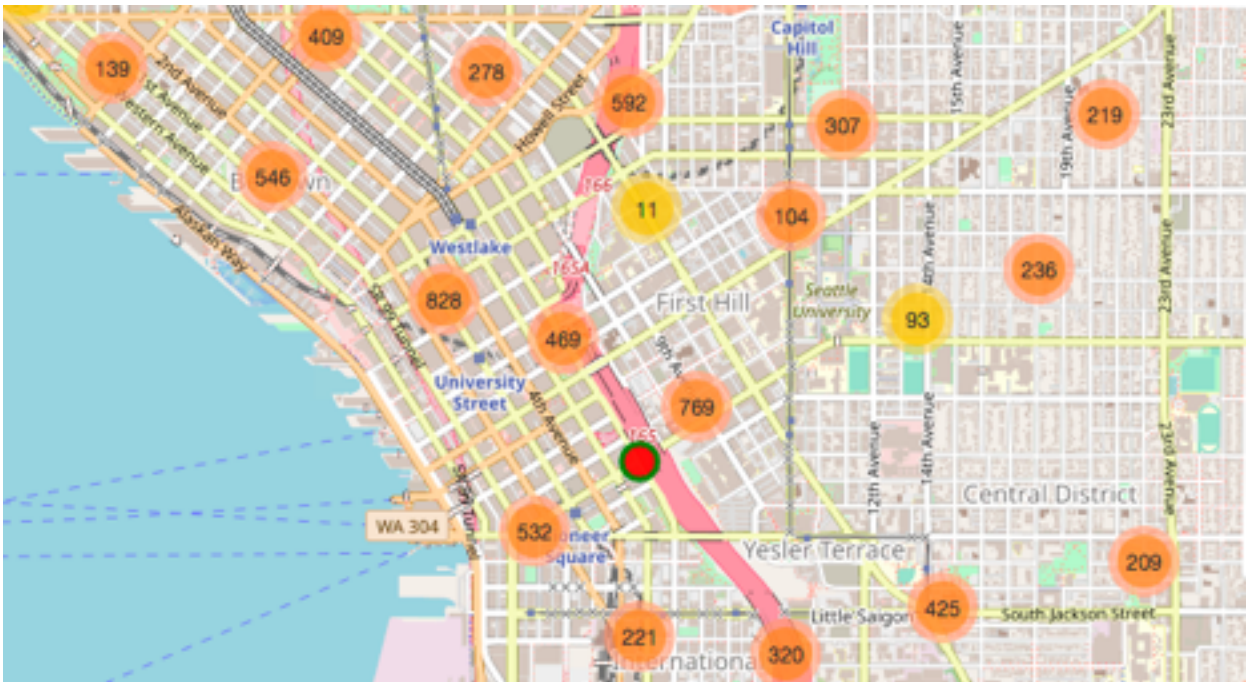

Figure 2a. View of the Intersection in the Downtown Area

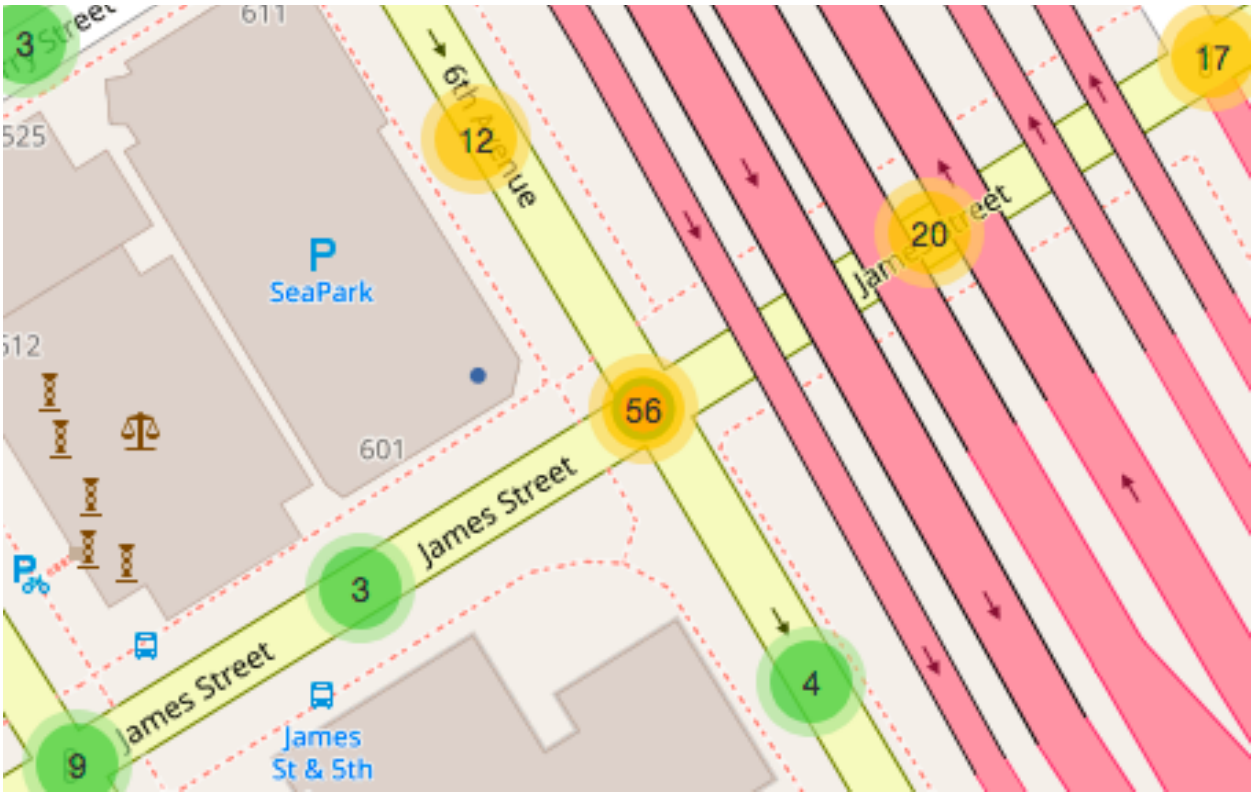

Figure 2b. A Close-Up of the Intersection

Accidents involving speeding were also observed on a city map (Figure 3). The markers are differentiated by color (Blue - No Injuries, Red - Injuries) because we are particularly interested in minimizing severe, injury-causing accidents. While no patterns emerged for differentiating severe accidents from others when speeding was involved, there was a high concentration of speed-related accidents along and near WA-509 near Boeing Field (Figure 4).

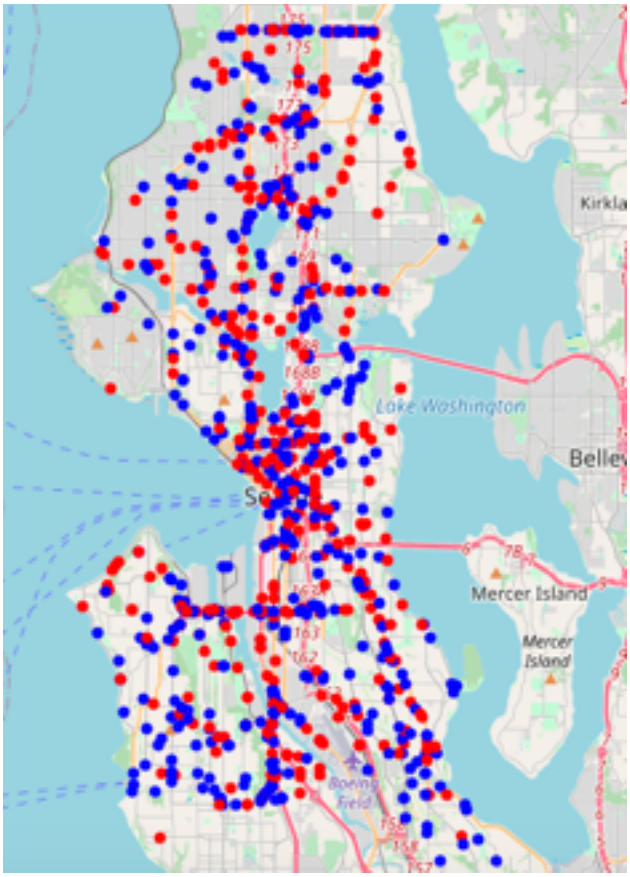

Figure 3. Speed-Related Accidents City-Wide

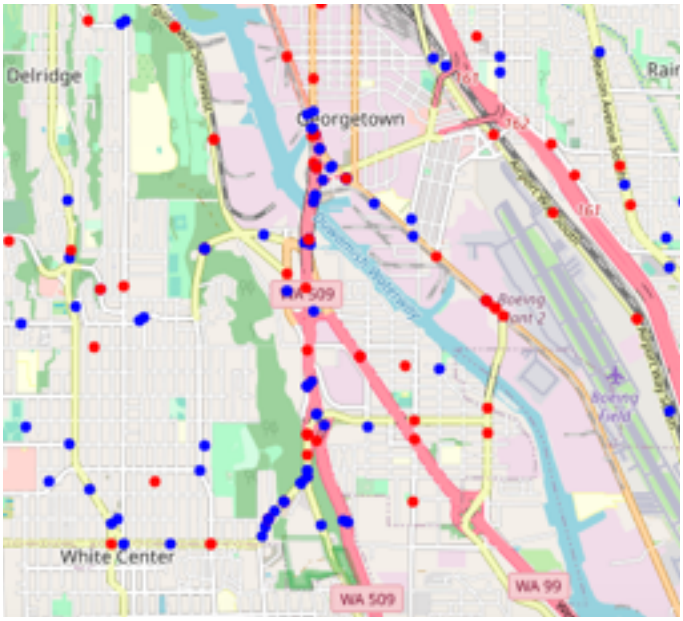

Figure 4. Speed-Related Accidents Along and Near WA-509

Next, we sought to create bar charts to help visualize minor and severe accidents in various weather conditions. Because of Seattle's reputation as such a rainy city with an average of 152 days of precipitation annually, we were motivated to determine what effect weather conditions had on the total number of accidents and the proportion of severe accidents. While the data set reported 10 different weather types, we only analyzed the top 5 different weather types as they represented more than 99% of accidents.

Given the resulting wet road conditions, we might reasonably assume that rain and snow would be associated with a high number of traffic accidents. However, the data indicates that, if anything, car accidents are more likely to happen in clear, sunny conditions. Furthermore, there is no statistically significant difference in the proportion of severe accidents on clear days and the proportion of severe accidents on wet weather days.

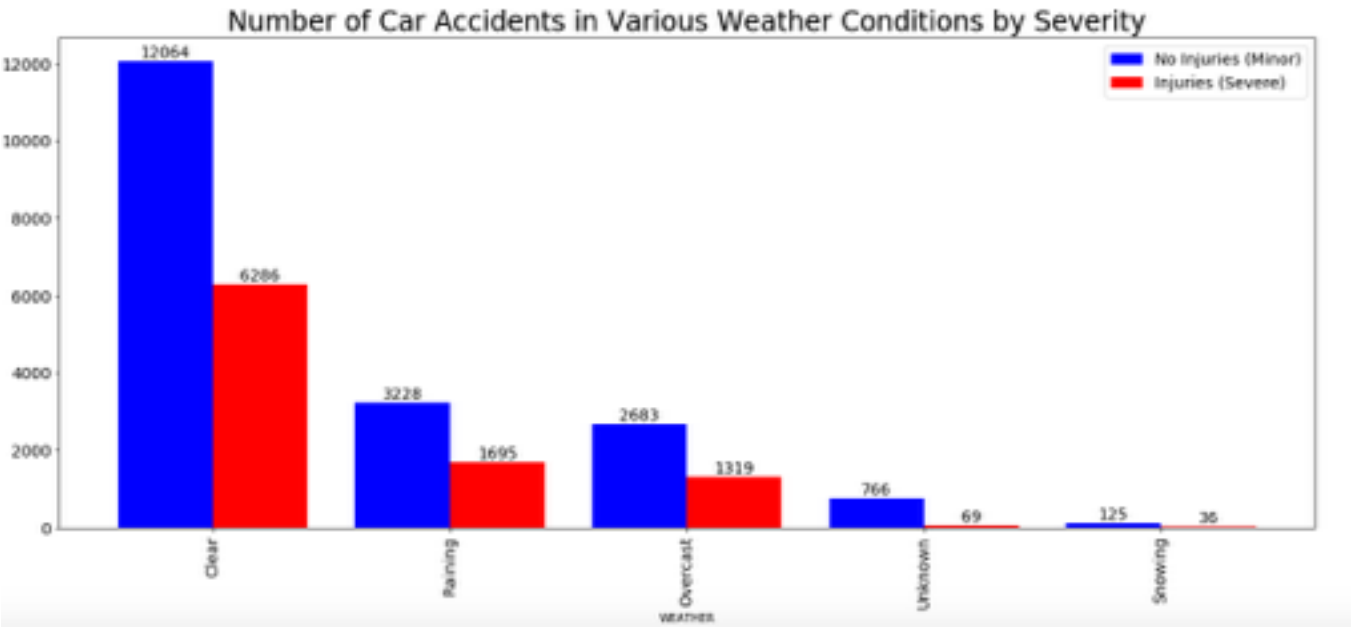

Figure 5. Number of Car Accidents in Various Weather Conditions by Severity

The percentage of severe accidents in clear weather was 34.26%, while the percentage of severe accidents in rainy and snowy weather was 34.43% and 22.36%, respectively.

A similar analysis and visualization was performed on road conditions data with nearly identical results.

4. Predictive Modeling

In order to build our model, we used the following features from the dataset: Location (X- and Y-Coordinates), Weather Conditions (One Hot Encoded), Speeding, and Under the Influence. We removed rows with no weather data, and replaced null values in the Speeding and Under the Influence columns the a 'no' value.

The target variable was Severity Code, since we want to use the features to predict the severity of a given accident. The dataset is unbalanced with only one-third (1/3) of the accidents classified as severe (injury-related). We balanced this data by up-sampling the minority class before implementing our Logistic Regression and K-Nearest Neighbors models.

We changed the values of the features in the dataset to numerical values. We then separated the feature variables from the target variable and split our data into training and testing sets.

We began with an Support Vector Machine (SVM) model with a radial basis function (rbf) kernel and balancing the weighting of the dataset target class by penalizing the model for failing to correctly identify the minority class. The model performed modestly with a F1-Score of 0.568 and Jaccard-Score of 0.560. Even with balanced weightings, it performed quite poorly with respect to the minority class. This shouldn't be terribly surprising as we expect SVM to excel with "wide" data as opposed to our situation here, where we a vary large number of accidents with relatively few features.

Next, we ran a random forest classifier model. Again, there was no need to balance the data yet, as these models tend to handle unbalanced data quite well. This model outperformed the SVM, obtaining an F1-Score of 0.600 and Jaccard-Score of 0.610.

Before running the K-Nearest Neighbors and Logistic Regression models, we up-sampled the minority class data.

Next, we sought to determine what value for k would work best for our KNN model, and we determined that k = 1 was that value (Figure 6).

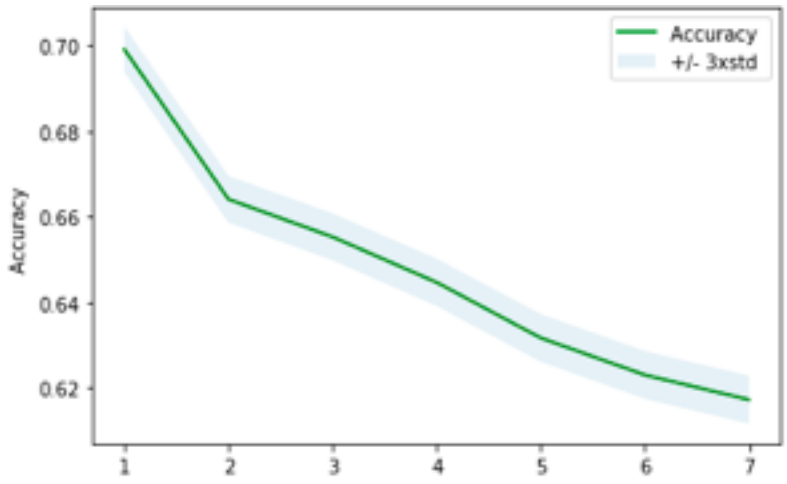

Figure 6. Accuracy of KNN Model for k Values from 1 to 7

This model performed much better than the others, with a F1-Score of 0.698 and Jaccard-Score of 0.699. However, there appears to be some degree of over-fitting, as the prediction accuracy of training data significantly exceeded that of the test data. Even still, this model is far superior to the others even in classifying minority class data.

Our final model was a Logistic Regression model, for which we used the up-sampled dataset again. This model performed most poorly with an F1-Score of 0.517 and Jaccard-Score of 0.521. However, it is worth noting that this model performed relatively well with minority class data, second only to the KNN model.

A summary of the evaluation metrics associated with each model can be found in Figure 7.

ALGORITHM	F1 SCORE	JACCARD SCORE
Logistic	0.516749599	0.520860666
SVM	0.567583865	0.559740715
Decision Tree	0.599676821	0.610721794
KNN	0.697891257	0.699160325

Figure 7. Evaluation Summaries for the Four Models.

5. Conclusion

In this study, we analyzed data with a wide range of features to identify possible avenues for reducing traffic accidents, especially severe ones. We analyzed intersection data to identify locations with a high number of accidents and proportion of severe accidents. We also mapped location data to observe areas with high numbers of accidents and areas where speeding was frequently a factor. We also analyzed weather and road condition data to determine if wet conditions positively correlated with increased accident numbers or severity. Lastly, we constructed a useful model for predicting an accident's severity with a view to minimizing the factors associated with severe accidents and the emergency response time when such accidents occur.

6. Future Direction and Recommendations

For future study and model development and improvement, the following features could be added to the dataset: 'Texting while Driving'(Y/N), 'Vehicle Data' (year, make, model), 'Insured Vehicle' (Y/N), and 'Major Event' (Y/N - based on whether a major event such as a professional football game occurred within a specified radius at the time of the accident).

Contrary to expectations, we found that wet weather and road conditions did not play a role in increasing the number nor severity of traffic accidents.

We propose that the James St. - 6th Ave. intersection be monitored and investigated further to determine an action plan to minimize the number and severity of accidents there.

We also propose that action be taken to decrease the number of accidents along the I-5 corridor such as the promotion of public transport use to lower the number of vehicles on the road. An alternative action would be a feasibility study aimed at increasing the number of traffic lanes for these troublesome routes.

Lastly, we suggest increasing surveillance along and near WA-509 near Boeing Field to lower the instances of speeding vehicles before they result in accidents.