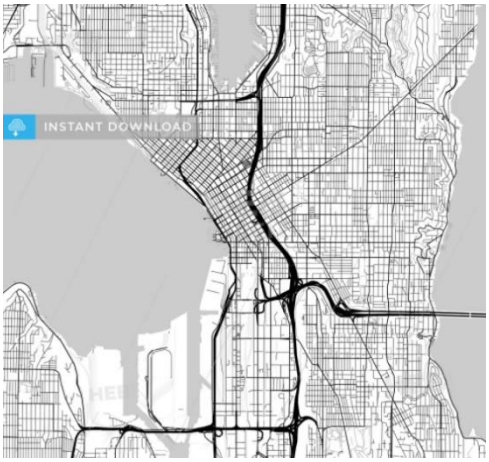
PREDICTING SEATTLE´S SEVERE COLLISIONS



RENÉ FERNANDO DUARTE

SEPTEMBER 2020

TABLE OF CONTENT

1. INTRODUCTION

1.1 PROBLEM AND DATA DESCRIPTION

Driving represents the most common transportation method of the modern societies, due to the large number of users there are many accidents using this transportation method that societies and governments all around the world need to consider and work on different aspects and rules to solve them.

With this project I will be able to analyze different conditions such as weather, light, location, among others, making a severe collision and leading to traffic, as well as showing all the vulnerable locations and type of accidents that mostly occur in Seattle. The societies and governments can use this information to create awareness and develop measurements to diminish these accidents. The findings will be achieved by machine learning algorithms that can predict the dangerous locations when the conditions present themselves using the historical data as reference.

All the data cleaning process will be done to fit the Data to Machine Learning Methods, such as Machine Learning Methods: K-nearest neighbor, Decision Tree, and Support vector machine. The locations will be drawn into a map using python Data Visualization, xls and Plateau.

2. EXPLORING THE DATA

2.1 DATA SOURCE

The historical data consists on a file provided by Coursera and has been compiled by the Seattle Police Department from 2004 until the present 2020. It has 38 columns and about 200,000 incidents(accidents) that demonstrates all the attributes of the accident; there are several conditions that are dependant to each Accident, as well as independent conditions in the Data. After analyzing the data, we can observe that Severitycode measures the damage of the collision and characteristics and will be used as damage reference throughout the study.

2.2 DATA CONTENT

Taking into consideration the almost 200,000 accidents showed per each row, the data shows under each column a different attribute that helps to describe the location, incident, information about the impact, number of vehicles, number of people, weather, light and road conditions among others.

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN

Table 1. Original Dataframe, first 4 rows and first 13 columns

3. METHODOLOGY

3.1 DATA CLEANING

For the Columns Weather, Road condition, and Light Condition showing unknown data will be replaced by Other. The rows with data showing "Blank" status will be deleted, the rows with “Unknown” will be added to “Other”.

The small categories in each column showing small counts will not be considered in most of the study, as the study wants to analyze the most influential/repetitive counts and will also be added to “Other” category.

All the NaN and Dummy values will be cleaned when converting Columns to Categories such as Weather Column to be transformed to 0s and 1s. The same with Road & Light conditions. All the columns without values will be eliminated.

At the end all the relevant data should show no string values and will be ready to be normalized.

The below example for column “ROADCOND” and it was applied to “WEATHER”, “JUNCTTYPE” & “LIGHTCOND”.

```
df_data_1['ROADCOND'].value_counts()
```

```
t[18]: Dry                124510
       Wet                 47474
       Unknown             15078
       Ice                  1209
       Snow/Slush           1004
       Other                 132
       Standing Water        115
       Sand/Mud/Dirt          75
       Oil                    64
       Name: ROADCOND, dtype: int64
```

```
Conditions1 = pd.DataFrame(df_data_1
Conditions1.head()
```

2]:

SEVERITYCODE	ROADCOND
2	Dry
1	Wet
1	Wet
1	Wet
2	Dry

Figure 1. Counting values for column “ROADCOND” and comparing the values to Severity code

WEATHER	ROADCOND	LIGHTCOND	ST_COLDESC	HITPARKEDCAR
Overcast	Wet	Daylight	Entering at angle	0
Raining	Wet	Daylight	Entering at angle	0
Clear	Dry	Daylight	Entering at angle	0
Raining	Wet	Daylight	Entering at angle	0
Clear	Dry	Daylight	Vehicle Strikes Pedalcyclist	0

Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind
0	0	0	1	0	0	0
0	0	0	0	0	1	0
1	0	0	0	0	0	0
0	0	0	0	0	1	0
1	0	0	0	0	0	0

Table 2. Transforming Column “ROADCOND” to 0s and 1s to be able to analyze the columns

3.2 MACHINE LEARNING MODELS

3.2.1 K-Nearest Neighbor

Once all the columns and row have been normalized (including the test & train data sets) a K = 10 is modeled in the KNN and a chart with the best achievable accuracy is created.

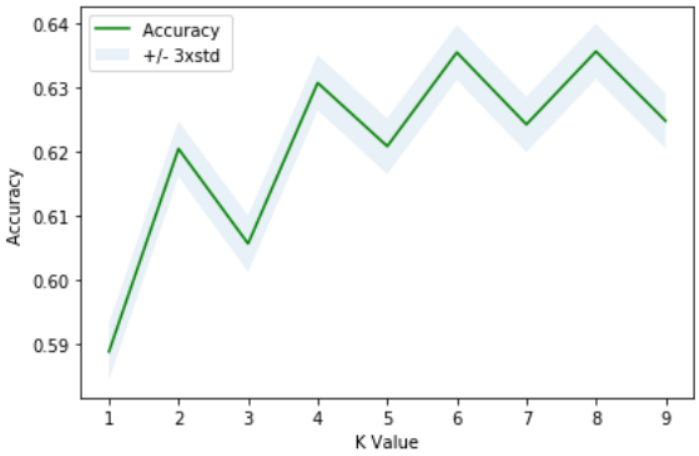


Figure 2. Finding best K Value with Accuracy

The best accuracy result for Train & Test sets are using K "8" values with a 0.7049 & 0.6390 accordingly using the model K – Nearest Neighbor.

When we use this model to analyze the weather, road & light conditions we have a 63.90% accuracy at predicting the severe accident from occurring in specific locations; the KNN needs additional validation method to measure the accuracy without overfitting data.

3.2.2 Decision Tree

To be able to compare and add validation to the accuracy of our model, a decision tree model is made using the same train and test sets.

```
predTree = accitree.predict(X_testset)

from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_testset, predTree))
```

```
DecisionTrees's Accuracy:  0.6112198994442974
```

The accuracy of the Decision Tree Model, when used to predict the severity of traffic collisions in Seattle, is 61.12%. The Decision Tree Model was used to correct the predictions made by taking multiple algorithms & conditions into account.

3.2.3. Support Vector Machine

```
from sklearn import svm
clf = svm.SVC(kernel='rbf')
clf.fit(X_train, y_train)
```

```
/opt/conda/envs/Python36/lib/python3.6/site-packages/sklearn/svm/base.py:19
nge from 'auto' to 'scale' in version 0.22 to account better for unscaled f
avoid this warning.
  "avoid this warning.", FutureWarning)
```

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=False, random_state=None,
  shrinking=True, tol=0.001, verbose=False)
```

```
print("Support Vector Accuracy: ", metrics.accuracy_score(y_test,yhat))
```

```
Support Vector Accuracy:  0.6248313090418354
```

Using the Support Vector Machine as a linear model to assure the severity of future accidents with a 62.48% of accuracy.

3.2.4 Applying Machine Learning models

Using the historical data and having all the latitudes and longitudes locations of all the collisions we can predict the places that will have a collision when the conditions like weather, road conditions, light, speeding, among others present. They will be presented in the Result section below.

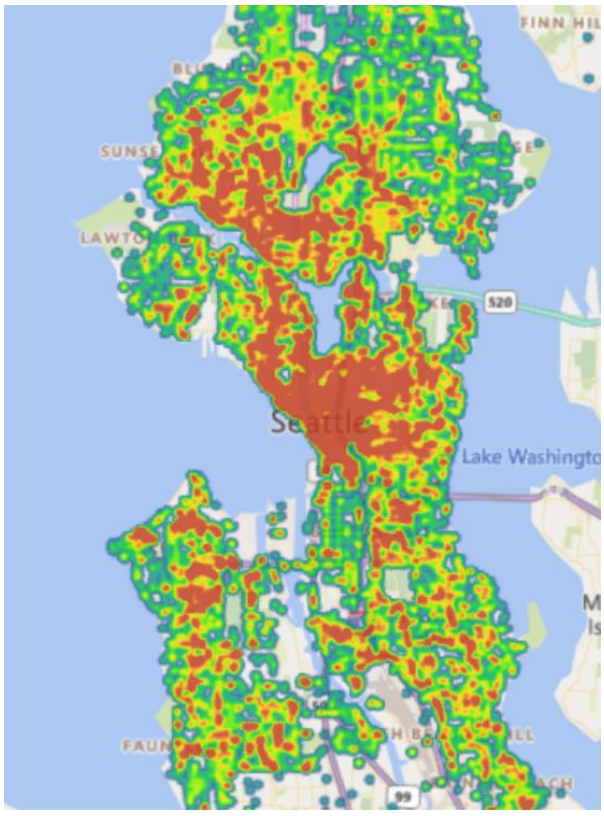


Figure 3. Map of all Accidents Seattle 2004 - 2020

4. RESULTS

4.1 MODELS

MODEL	ACCURACY
K-Nearest Neighbor	63.90 %
Decision Tree	61.12 %
Support Vendor Machine	62.48 %

Table 3. Machine Learning Model results

In order to predict the locations of possible severity collision in Seattle, the maps are created using the Support Vendor Machine results of 62.28% accuracy.

4.2 DATA ANALYSIS RESULTS

4.2.1 Seattle Total Accidents vs Traffic

YEAR	NUMBER OF COLLISIONS
2004	**11418**
2005	**13666**
2006	**13745**
2007	**12989**
2008	**12208**
2009	**11207**
2010	**10789**
2011	**10846**
2012	**10086**
2013	**10209**
2014	**11449**
2015	**9224**
2016	**6541**
2017	**5685**
2018	**4810**
2019	**4534**
2020	**802**
TOTAL	**160208**

Table 4. Yearly Collisions

The Collisions or Accidents have been considerably dropping since 2014, along with the traffic itself in Seattle. Both follow the same improvement pattern.

Figure 4. Seattle Collisions 2004 – 2020 Timeline

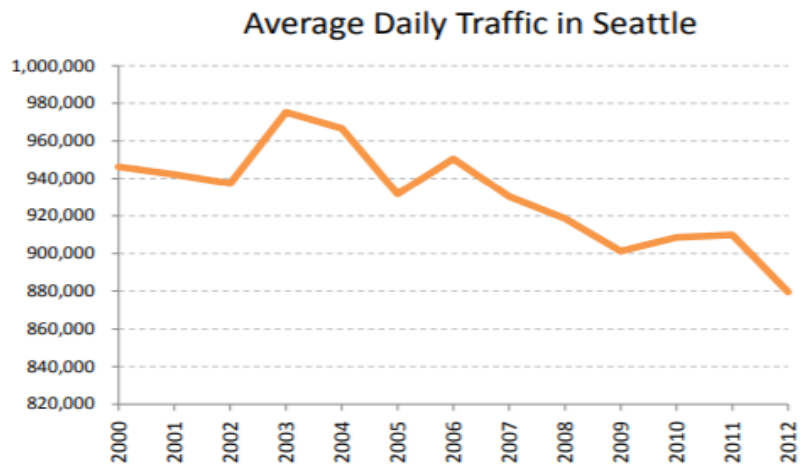


Figure 5. Seattle Traffic 2000 – 2012 Chart. Source: Seattle Police Report

4.2.2 Seattle Weather per Location Type of Accidents

When analyzing the effects of the Weather with the collisions based on each Location type, most of the accidents occurred at Clear weather conditions, however, the risk increases in Raining conditions at Intersections.

Weather /Location Accident Type

Address Type	Clear	Other	Overcast	Raining
Alley	432	172	98	89
Block	71117	16549	17694	20403
Intersection	39077	3644	9792	12512
Other	509	1139	130	141
Total Accidents	**111135**	**21504**	**27714**	**33145**

Table 5. Yearly Collisions

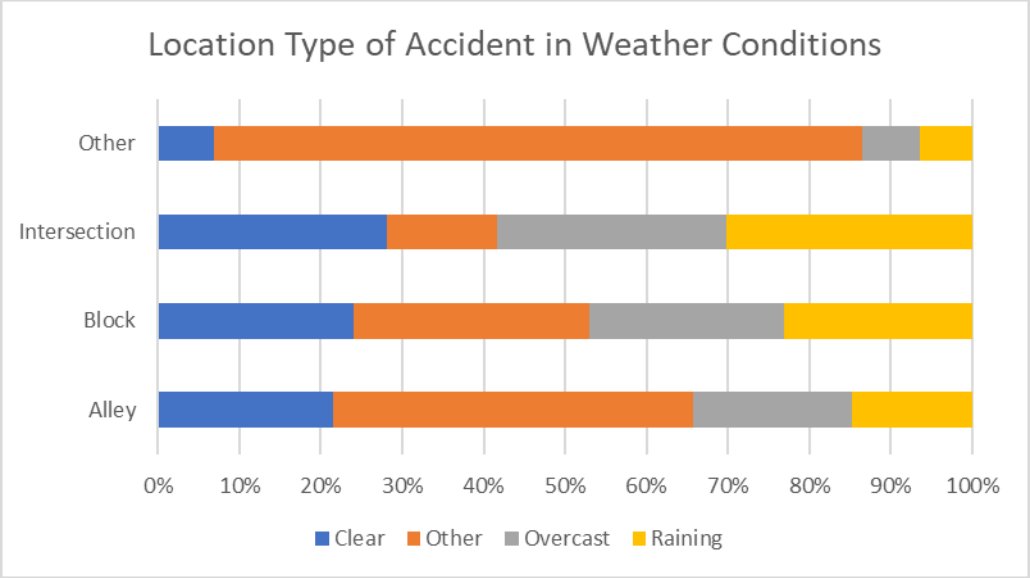


Figure 6. Location Type vs Weather Condition Collisions

4.2.3 Seattle Accident Severity under Influence

SEVERITY CODE	Anzahl von UNDERINFL	PERCENTAGE
1	5559	60.94%
2	3562	39.06%
From 252861 Accidents the following are for Alcohol/Drug Abuse:	**9121**	

Table 6. Severity Code under Influence Collisions

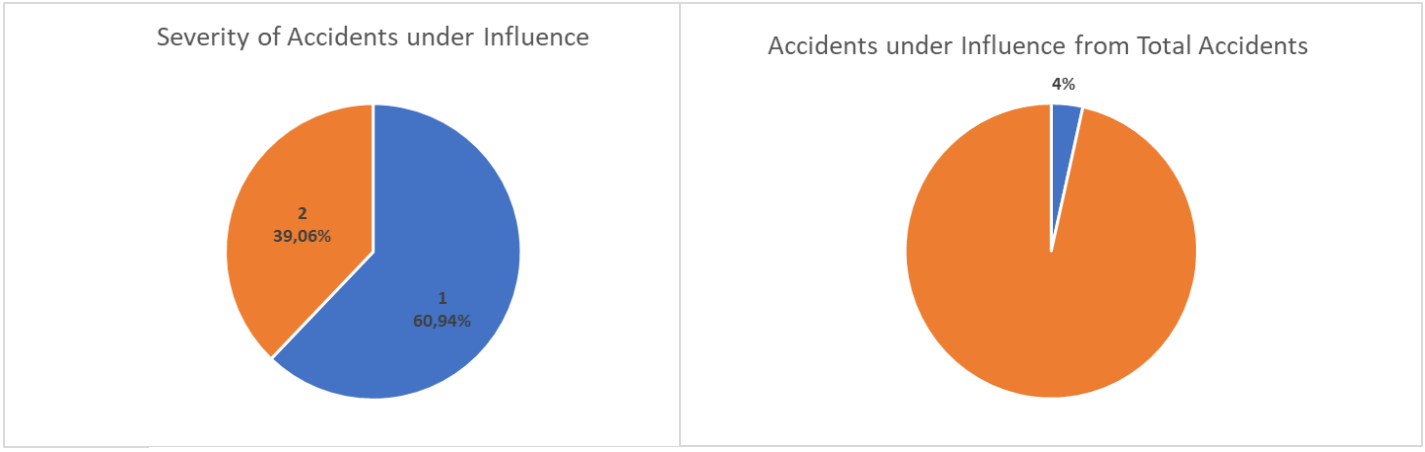


Figure 7. Severity of Accidents under Influence & Accidents under Influence from Total Accidents Charts

Most of the accidents under Drug/Alcohol influence caused a Severity 1 collision and only 4% of the total Accidents were caused by driving under influence.

4.2.3 Seattle Accidents per Light Conditions

We previously analyzed that most collisions occurred in clear weather, now we re-confirm that most accidents happened during Daylight.

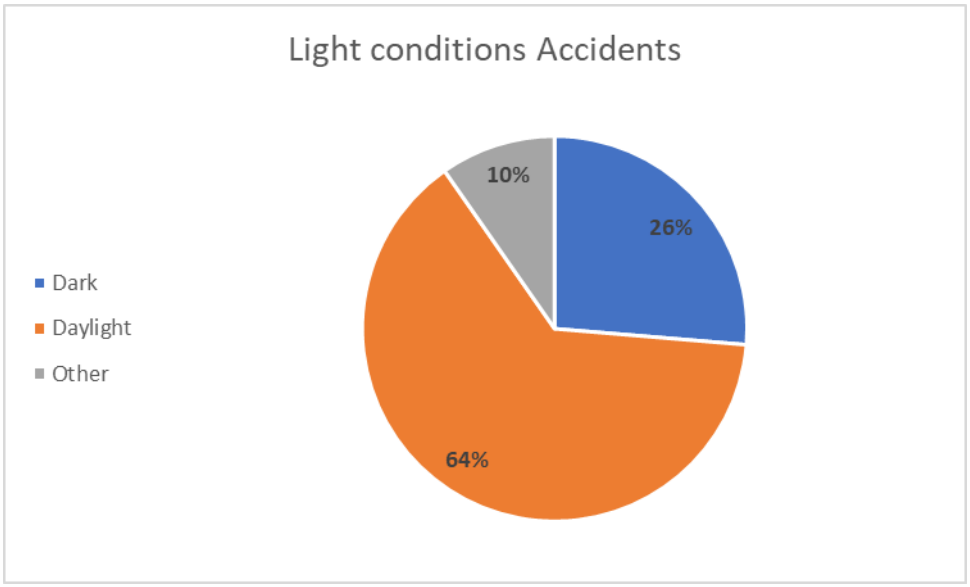


Figure 8. Collisions under Light Conditions

4.3 SEATTLE COLLISION MAPS

SEATTLE HEATMAP WITH SEVERITY 1

SEATTLE HEATMAP WITH SEVERITY 2

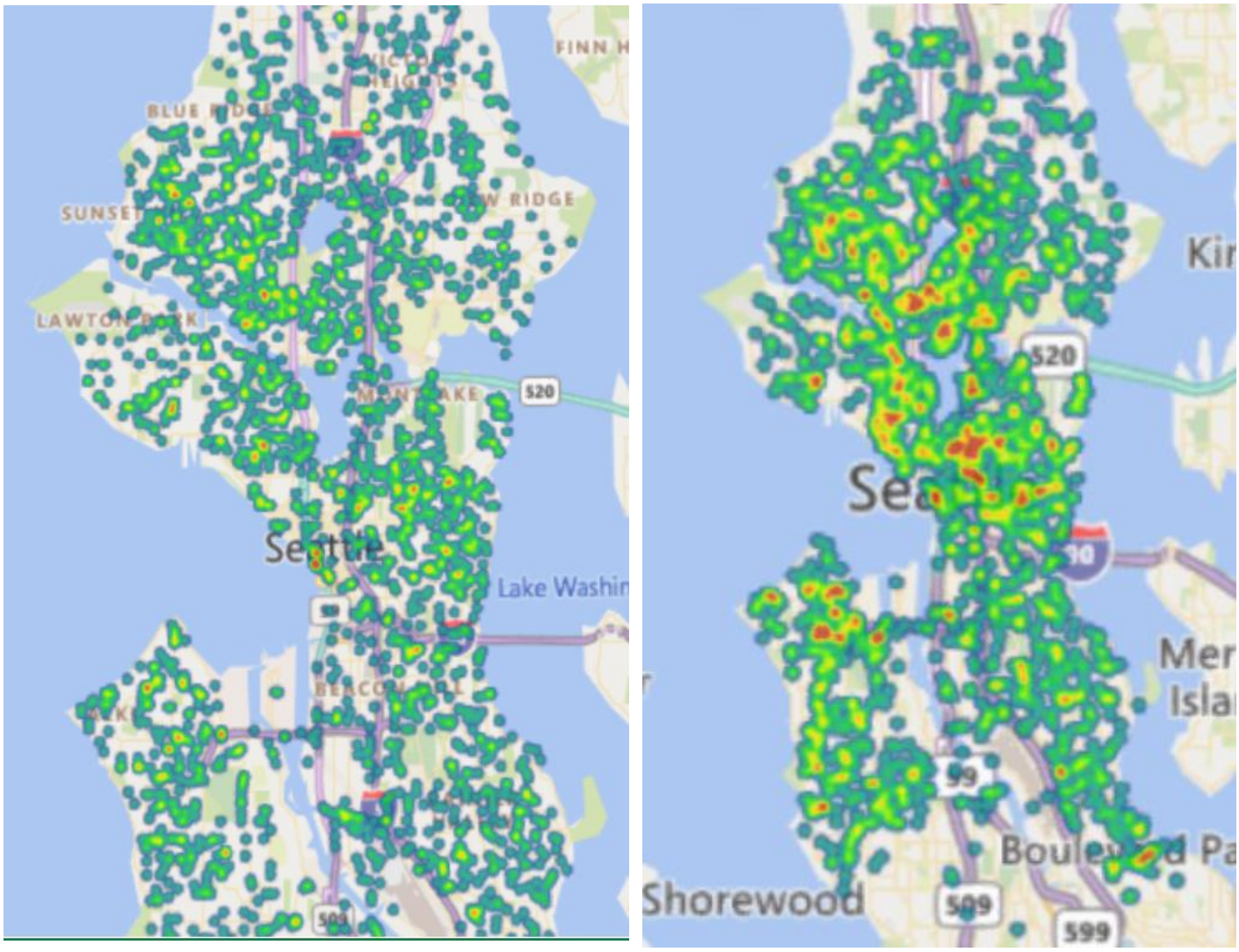


Figure 9. Seattle Severity 1 & 2 collisions map

A historic comparison map between Severity collisions type 1 and type 2 show us that the areas where the collisions type 2 occur are more repetitive.

There are a total of 136,485 collisions of severity type 1 and 58,188 type 2.

SEVERITY 2 WITH ROAD CONDITIONS WET, SNOW, ICE & DAYLIGHT

SEVERITY 2 WITH ROAD CONDITIONS WET, SNOW, ICE & DARK

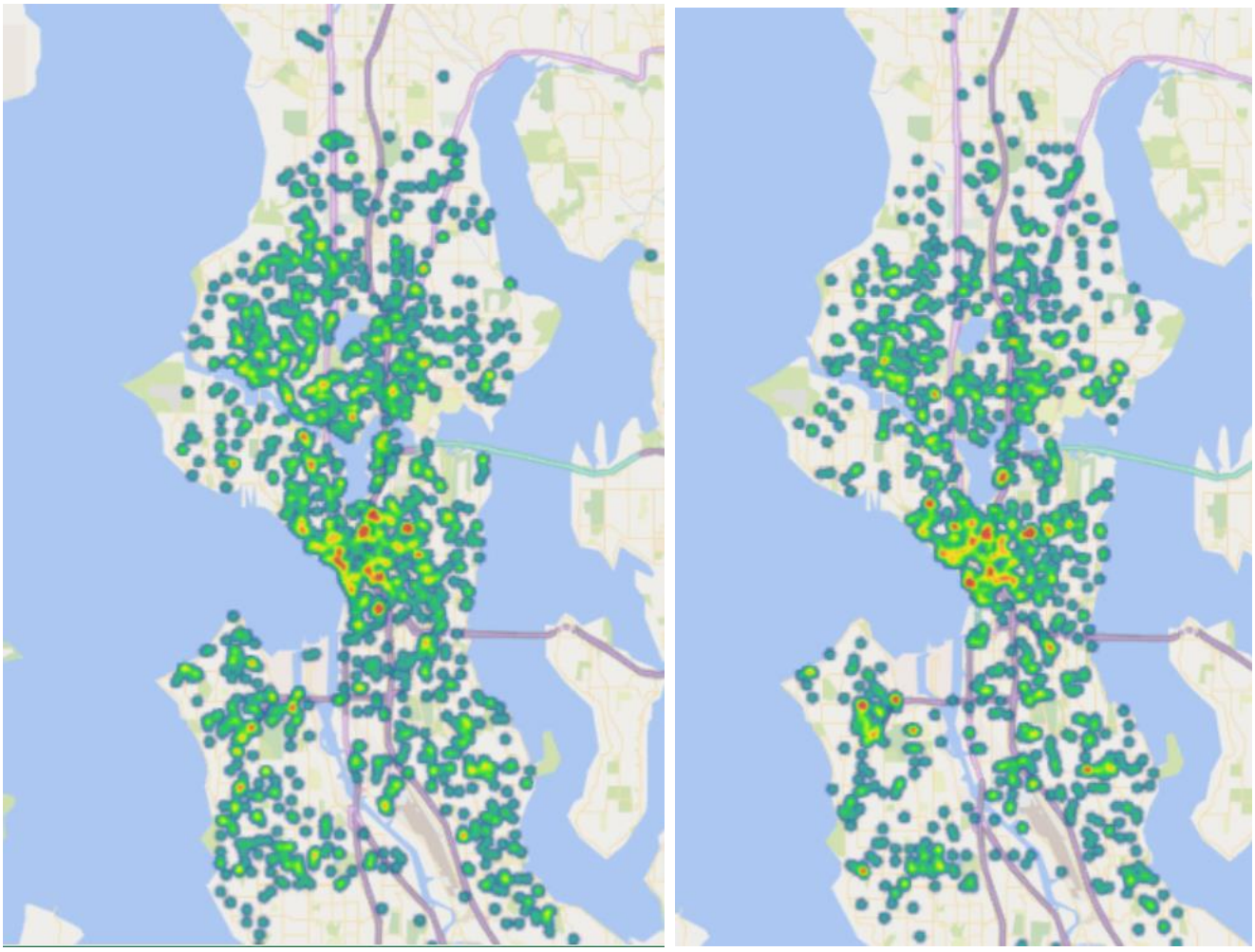


Figure 10. Seattle Severity 2 collisions with road conditions wet, snow and ice in daylight & at dark with and without light map

Although there are differences in the collisions with severity 2, road conditions being wet, with snow or ice, with or without light there is not a major difference that can clearly indicate that not having lights on the streets under these conditions is the main cause for the collisions.

The road condition does not contribute directly to a severity 2 collision.

SEVERITY 2 WITH RAINING, OUTCAST, ROAD CONDITIONS WET, SNOW, ICE & DARK WITH STREET LIGHTS & ILUMINATION

SEVERITY 2 WITH RAINING, OUTCAST, ROAD CONDITIONS WET, SNOW, ICE & DARK NO STREET LIGHTS & STREET LIGHTS OFF

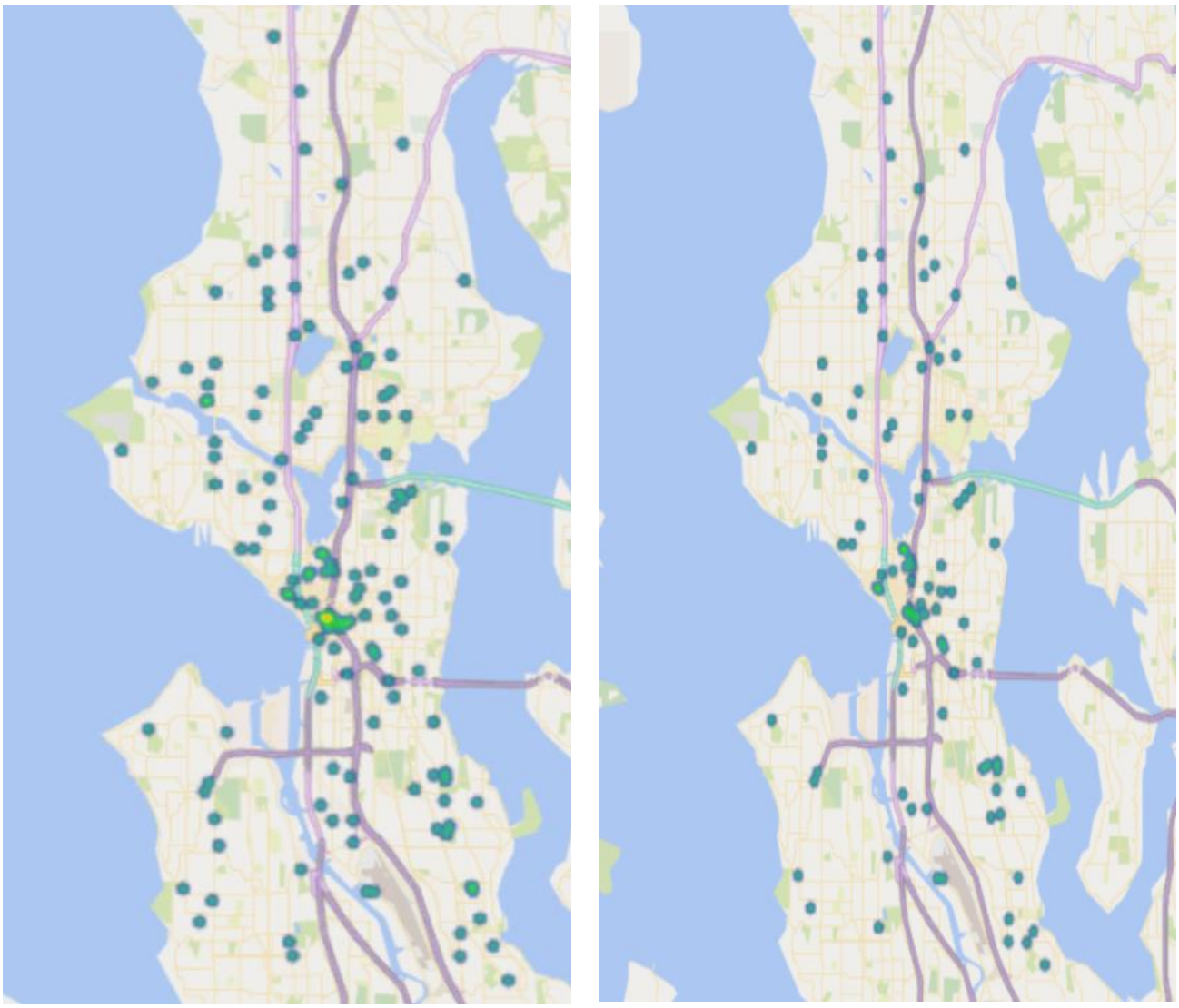


Figure 11. Seattle Severity 2 collisions with weather conditions raining & outcast, road conditions wet, snow and ice at dark without light map

Considering the quantity of the sample, the amount of collisions under the weather conditions raining and outcast (both are the weather conditions with most accidents) plus the road conditions wet, snow or ice, the light conditions did not affect the collisions, as there are more collisions during the day or with proper lighting than without.

UNDERINFLUENCE ACCIDENTS SEVERITY 1 & 2

SPEEDING ACCIDENTS SEVERITY 1 & 2

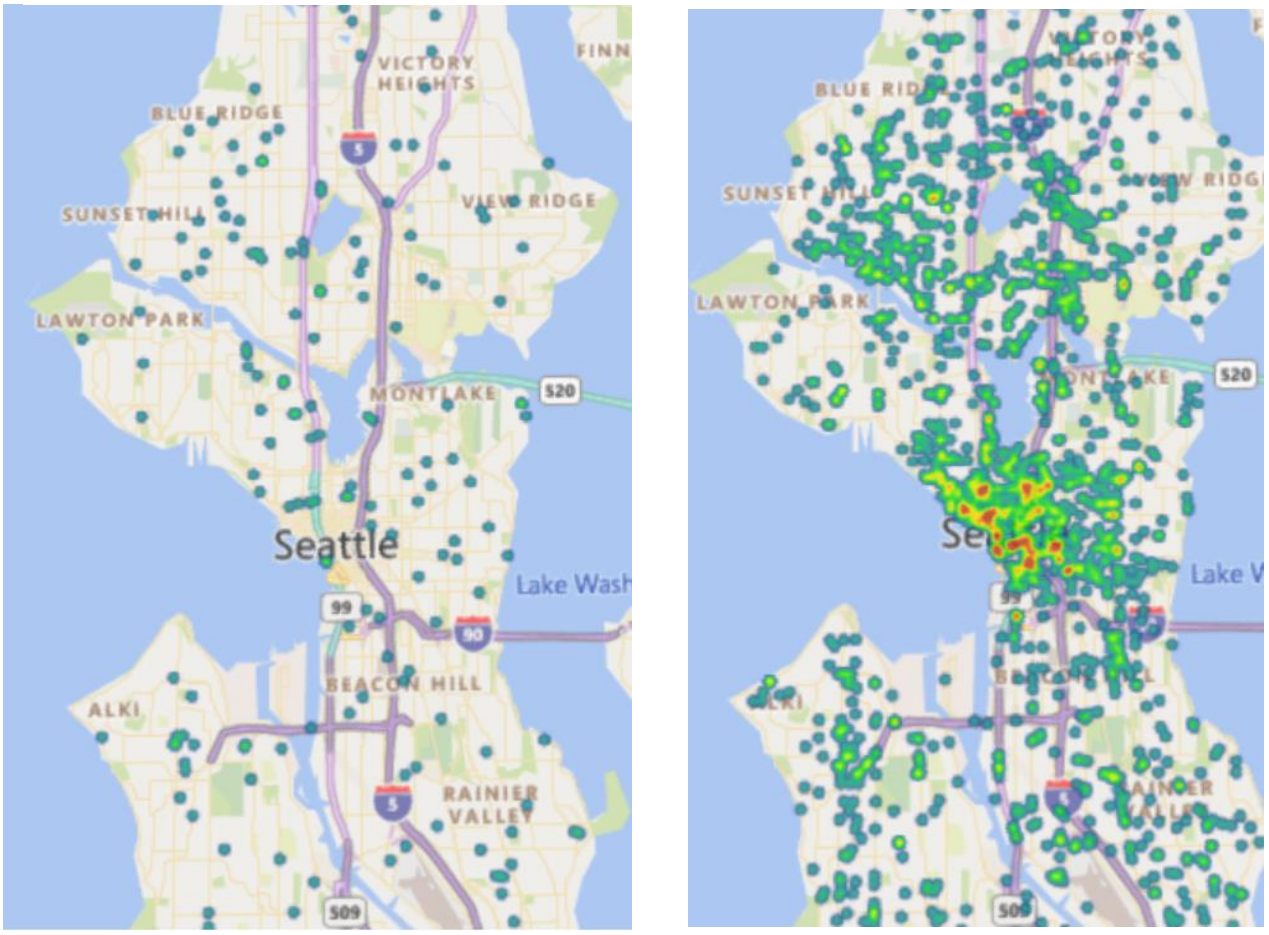


Figure 11. Seattle Severity collisions 1 & 2 under influence accidents & Speeding Accidents map

Although we previously discarded the collisions being caused by under influenced drivers, it is brought as a comparison for speeding. The results clearly show that most of the collisions were caused by speeding and independent of any weather, road or light conditions.

4.4 PREDICTION OF SEATTLE COLLISIONS

Using historical information plus the algorithms of machine learning using as base the SVM with 62.42% as reference, we can predict the places that under certain conditions will lead to collisions.

PREDICTING SEVERE COLLISIONS 1 SPEEDING LOCATIONS

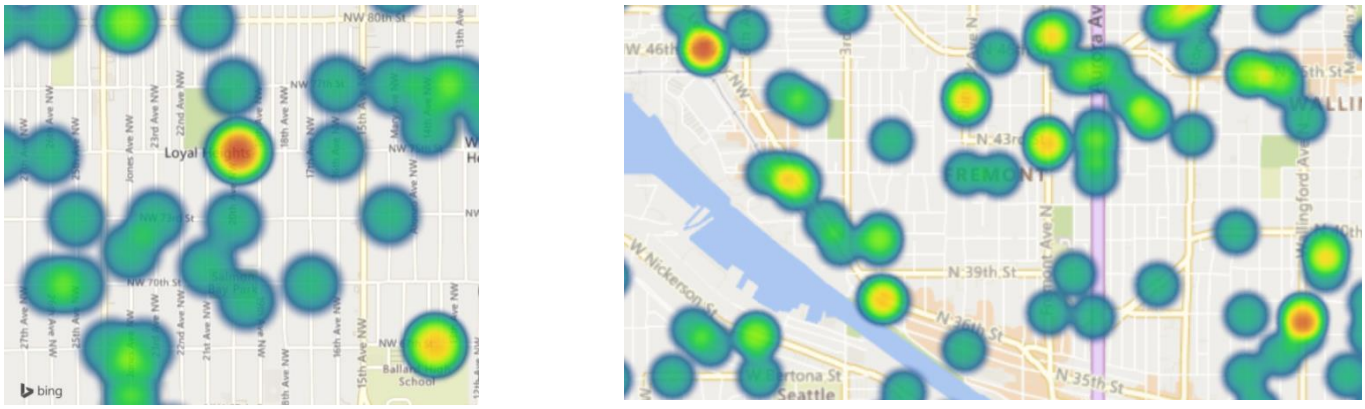


Figure 12. Prediction of Severe Collisions 1 location for Speeding map

The above locations have a higher tendency of severe collisions 1 when speeding conditions occur.

PREDICTING SEVERE COLLISIONS 2 SPEEDING LOCATIONS

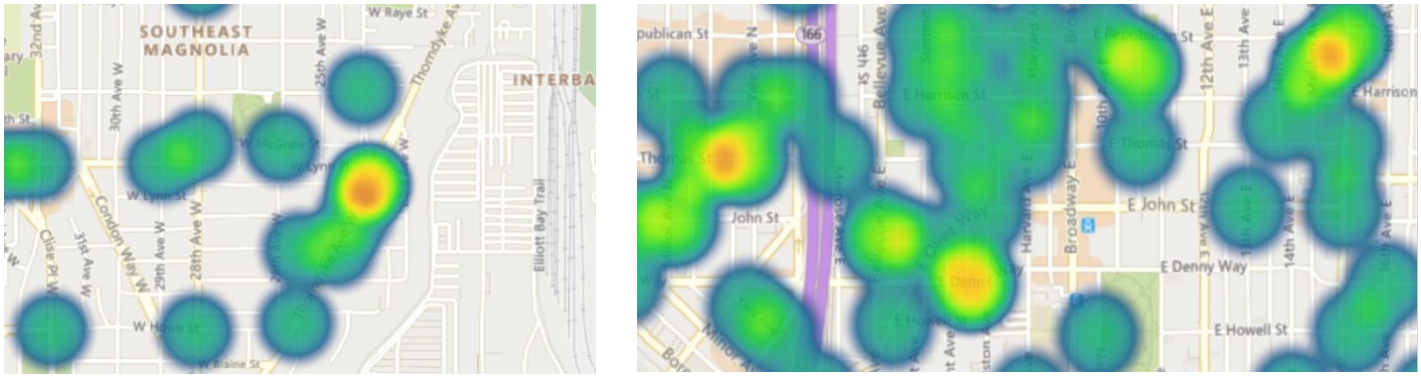


Figure 13. Prediction of Severe Collisions 2 location for Speeding map

The above locations have a higher tendency of severe collisions 2 when speeding conditions occur.

PREDICTING UNDERINFLUENCE COLLISIONS SEVERITY 1 & 2

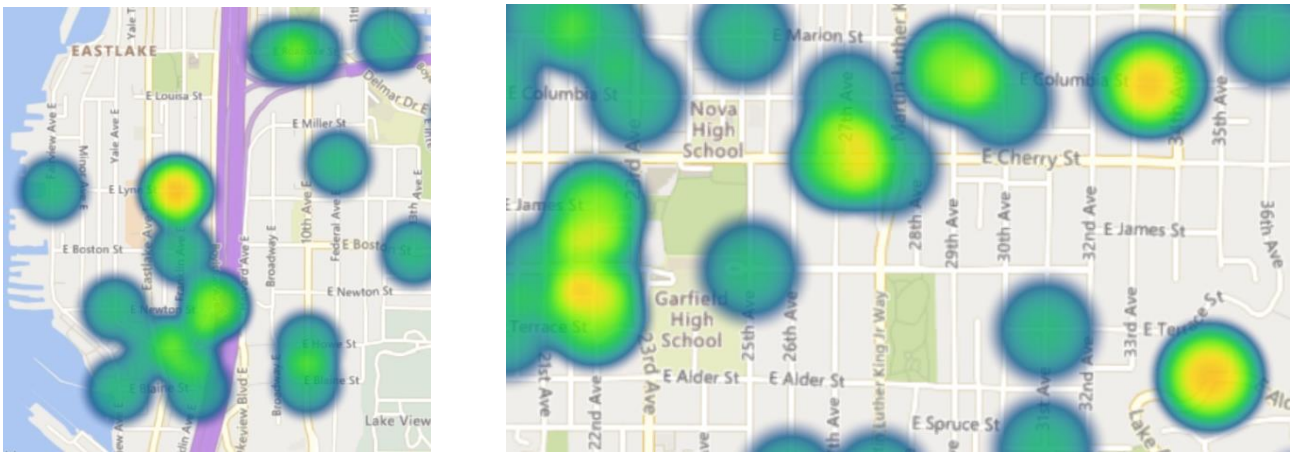


Figure 14. Prediction of Under Influence Collisions 1 & 2 location for Speeding map

PREDICTING COLLISIONS SEVERITY 2 WITH RAINING, OUTCAST, ROAD CONDITIONS WET, SNOW, ICE & DARK NO STREETS LIGHTS & MISSING ILUMINATION LOCATIONS

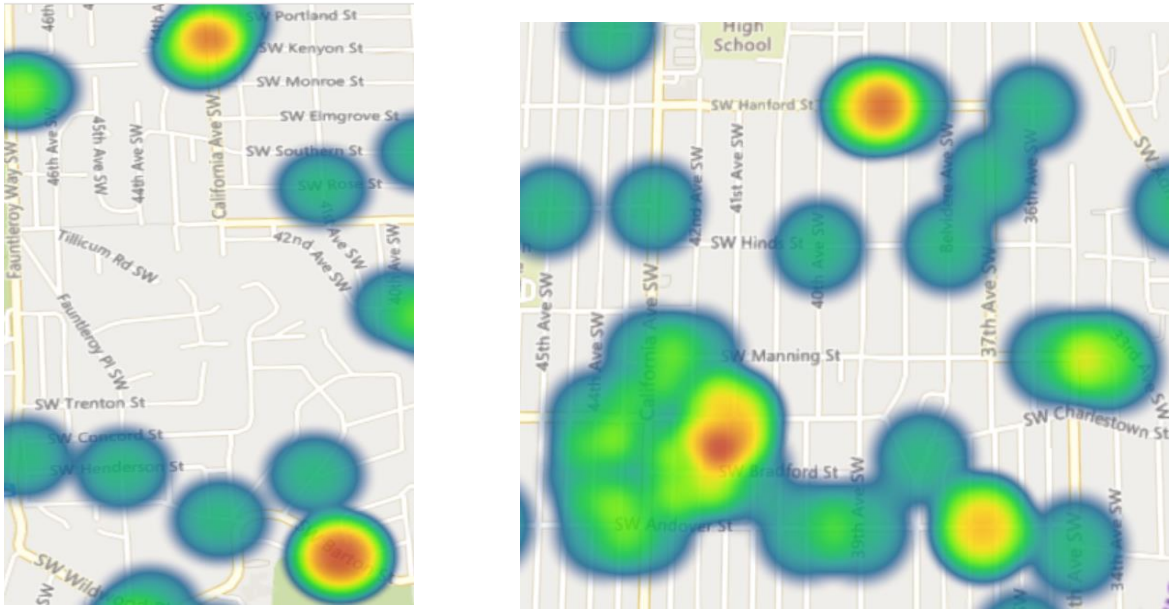


Figure 15. Prediction of Severity Collisions 2 locations when raining, outcast weather, wet , snow or Ice road conditions at dark without street lights and missing illumination map

PREDICTING COLLISIONS SEVERITY 1 RECURRENT LOCATION

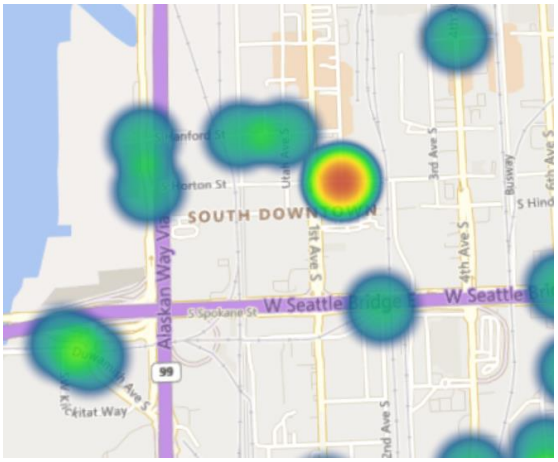


Figure 16. Prediction of Severity Collisions 1 locations map

PREDICTING COLLISIONS SEVERITY 2 RECURRENT LOCATIONS

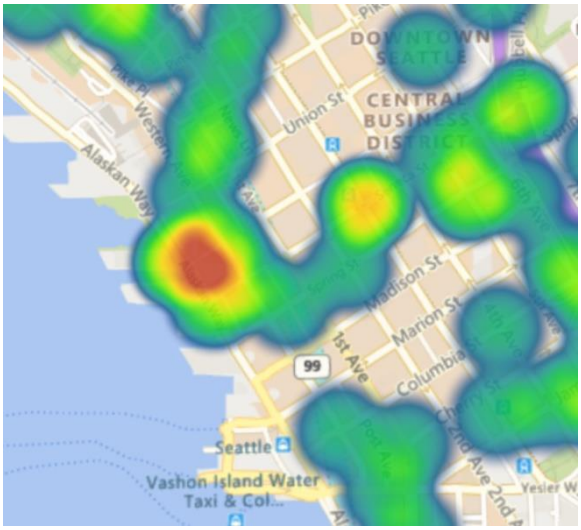


Figure 17. Prediction of Severity Collisions 2 locations map

5. DISCUSSION

The study does not intend to go into specific details about developing a live software or tool to track traffic and constantly predict new collision locations.

Using this model and location prediction per condition, a private or public transit office can alert and notify the drivers and local radio stations to avoid collisions that would also lead to an eventual traffic.

6. CONCLUSIONS

The collision severity can be predicted with a precise accuracy, the external conditions such as weather, visibility, road conditions can also be tracked and used to predict concurrent accident locations and traffic, however, as seen in this study, the external conditions do not influence or are not the main cause of the collisions.

Speeding is the most common cause of the severe collisions.

The use of the predictive locations based on each condition or type can be reinforced by the society or local government to prevent and improve the facilities, signs and lights on the streets.

A live-fed system is required to create an application or alert to inform all the drivers about the possible collisions when the conditions are present.

The significant reduction of collisions since 2004 has led to a major reduction of traffic in Seattle.