Predicting the severity of accidents in Seattle

Leyla Yunis

10 September 2020

1. Introduction

1.1. Background

Currently, the vast majority of countries population have one or more vehicles as means of transportation. This leads to more vehicles on the streets, and more accidents. In countries with large population, the number of vehicles, and collisions, are a common thing. For the population, it is important to avoid collisions at all costs, since they could end up badly hurt. But the probability of being involved in a collision is unknown for people. It is necessary to know the probability of being involved in a collision and the severity of it, so people can take action and drive more carefully or try to avoid it in high probability cases.

1.2. Problem

Data that might contribute to determining the probability and severity of a collision might include weather, specific streets, date, time, alcohol or drugs consumption, condition of the road, amount of light on the street, speed, and more factors that describe the general condition in which the collision took place. This project aims to the severity of a potential collision based on these data.

1.3. Interest

People that live in Seattle would be very interested in accurate prediction of collision probability and severity of it, for safety reasons. Others interested could be the mayor of the city and the government, so they could create safety

legislations based on this, improve roads, or make awareness campaigns that aims to reduce accidents.

2. Data

2.1. Data Source

The data on collisions and severities can be found in Seattle GeoData dataset here. This dataset is complete, and it has information from 2004 to the present, recorded by Seattle Department of Transport (SDOT). The dataset is in the form of a comma separated values file (.csv) where each row represents a collision. Each collision has a severity code label that allows for supervised machine learning.

2.2. Data Cleaning

Given the characteristics of the dataset, a significant amount of data preparation was required.

Dataset contains several missing values, where key variables are absent. This is the case of several columns. For example, column PEDROWNOTGRNT, which indicates if pedestrian right of way was granted or not, has 97.6% of its values missing. Similar with columns EXCEPTRSNDESC, SPEEDING, INATTENTIONIND, INTKEY, and EXCEPTRSNCODE, which all have more than 50% of missing values. Given this, I didn't consider these columns for the analysis and I deleted them from the dataset.

Dataset contains unnecessary or redundant variables that replicate information from other variables. OBJECTID, SEVERITYDESC and INCKEY are examples of this variables. I deleted these columns from the dataset because they did not give any valuable information.

After dropping the selected fields, there are still some missing values, where key variables are absent. With further inspection, 3.68% of all rows contains missing values. Since this is a small percentage, I decided to remove all these rows. This way, the dataset now contains no missing values.

The dataset only has two labels for the SEVERITYCODE field: 1 and 2, indicating property damage and injury, respectively. Collisions resulting in property damage are more than twice as common than collisions resulting in injuries, which we can see in the Figure below.

```
1    130634
2     56870
Name: SEVERITYCODE, dtype: int64
```

Figure 1: Unbalanced dataset.

With an unbalanced dataset, results will be biased and won't give correct predictions. So, I balanced the dataset using the resample method provided by

```
2    56870
1    56870
Name: SEVERITYCODE, dtype: int64
```

the **sklearn** library. After this stage, the dataset has the same number of rows with SEVERITYCODE 1 and 2 as we can see in the Figure below.

Figure 2: Balanced dataset.

Dataset also contains several categorical fields, such as COLLISIONTYPE, WEATHER, ROADCOND, LIGTHCOND, and ADDRTYPE, which are not suitable for performing quantitative analysis. So, it's necessary to recast them to an appropriate datatype. In this case, I chose to use a One-Hot Encoding technique to create a binary field for each category in a variable. This way we have several more fields, but it is easier to interpret the results.

2.3. Feature Selection

The feature selection was done in between the data cleaning, by removing redundant fields or fields that don't provide relevant information. Specifically, the features to consider are shown in the Table below, where the bottom 5 fields were categorical and recast into numerical fields as described in section 2.2. After this step, there were 45 features and 113740 samples in the data to perform the analysis.

Attribute	Description
SEVERITYCODE	Code that corresponds to the severity of the collision.
PERSONCOUNT	Number of people involved in the collision.
VEHCOUNT	Number of vehicles involved in the collision.
SDOT_COLCODE	Code given to the collision by SDOT.
ADDRTYPE	Collision address type.
COLLISIONTYPE	Collision type.
WEATHER	Description of weather conditions during the time of collision.
ROADCOND	Condition of the road during the collision.
LIGHTCOND	Light conditions during the collision.

Table 1: Feature Selection

3. Methodology

Data importing, data cleaning, exploratory analysis, modeling, and evaluation were all performed in IBM Cloud Watson Studio Platform through a Jupyter Notebook created within a project.

3.1. Exploratory Data Analysis

The matplotlib library package was used to visually inspect the data. Seven plots were created to compare different fields to the target variable, SEVERITYCODE, to see how the fields were related.

3.1.1. Collisions by Address Type

From the data we can clearly see that if a collision happened in a block, it's more likely to be a type 1 severity code, meaning, only property was damaged. Instead, if a collision happened in an intersection is more likely to be a type 2 severity

code, meaning, there would be injuries. Data also shows that alley collisions are extremely rare. (Figure 3)

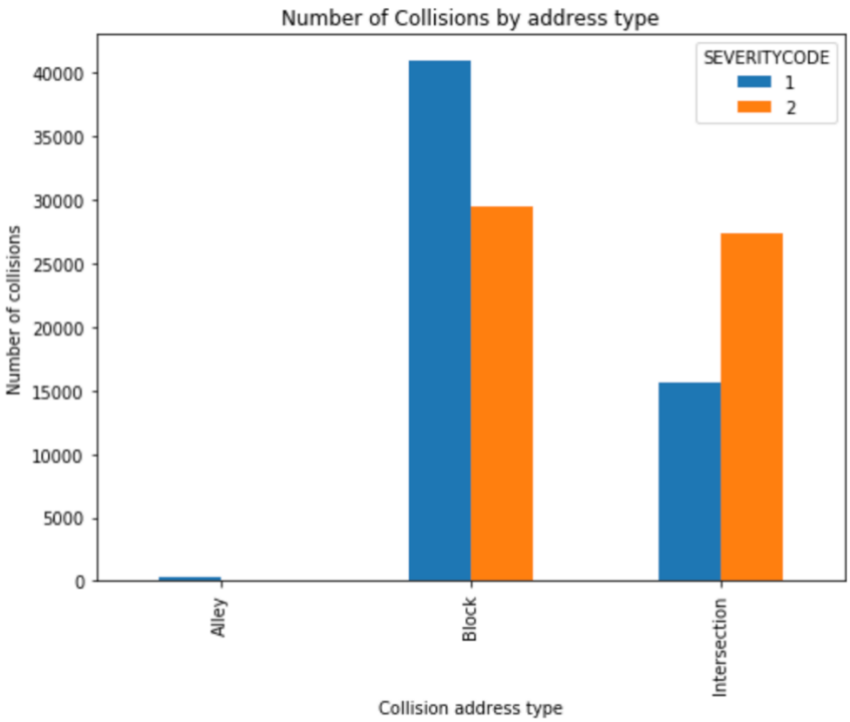

Figure 3: Collisions by Address Type

3.1.2. Collisions by Type

Data shows that collisions involving a parked car are one of the most common ones, but they result in mostly property damage. Other common collisions are caused by a car rear ending another one, and cars entering at an angle. These collisions often result in injuries. We can also see that most collisions involving bicycles or pedestrians results in injuries, and that a head on collision is not common in the dataset (Figure 4).

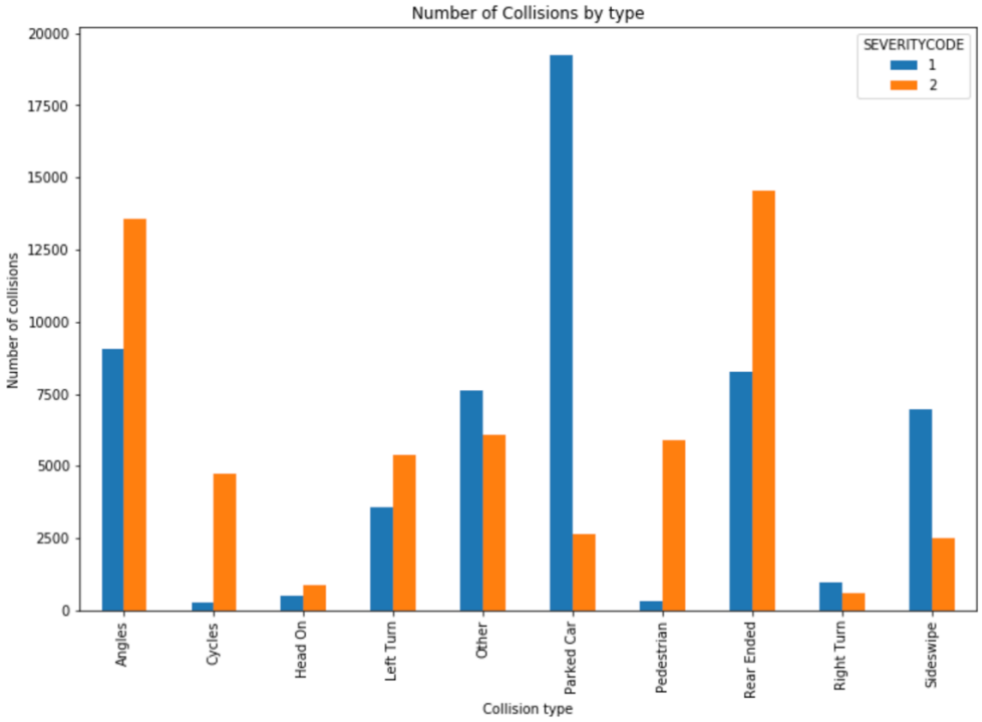

Figure 4: Collisions by Type

3.1.3. Collisions by Weather

From the data we can appreciate that most collisions happen when the weather is clear and not when the weather makes it harder to drive, like rain, snow, or fog. This could mean that people are less cautious when driving in a typically good weather. It is common to have collisions when there is an overcast or when it is raining. But it is not common to have collisions when there are other weather conditions (Figure 5).

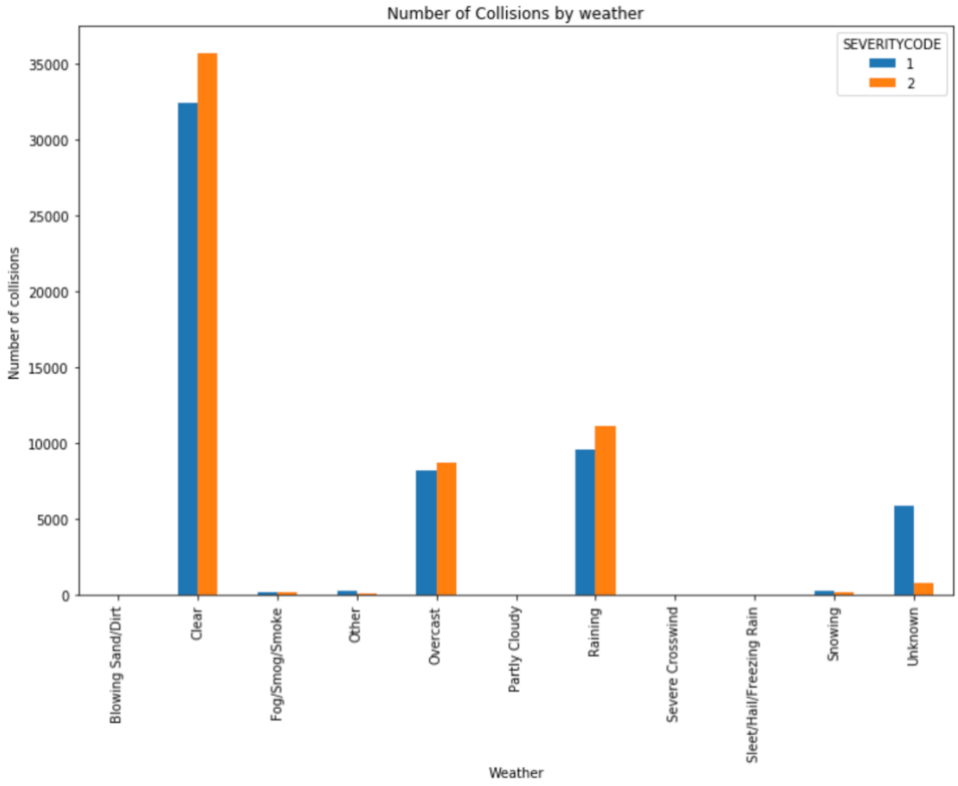

Figure 5: Collisions by Weather

3.1.4. Collisions by Road Conditions

The data shows that most collisions happen the most when the road is dry. This is contra intuitive, since wet or iced roads makes it harder to control a car and are typically known for causing collisions. There are multiple collisions registered where the road was wet, however, this is a much smaller number than when the road is dry (Figure 6).

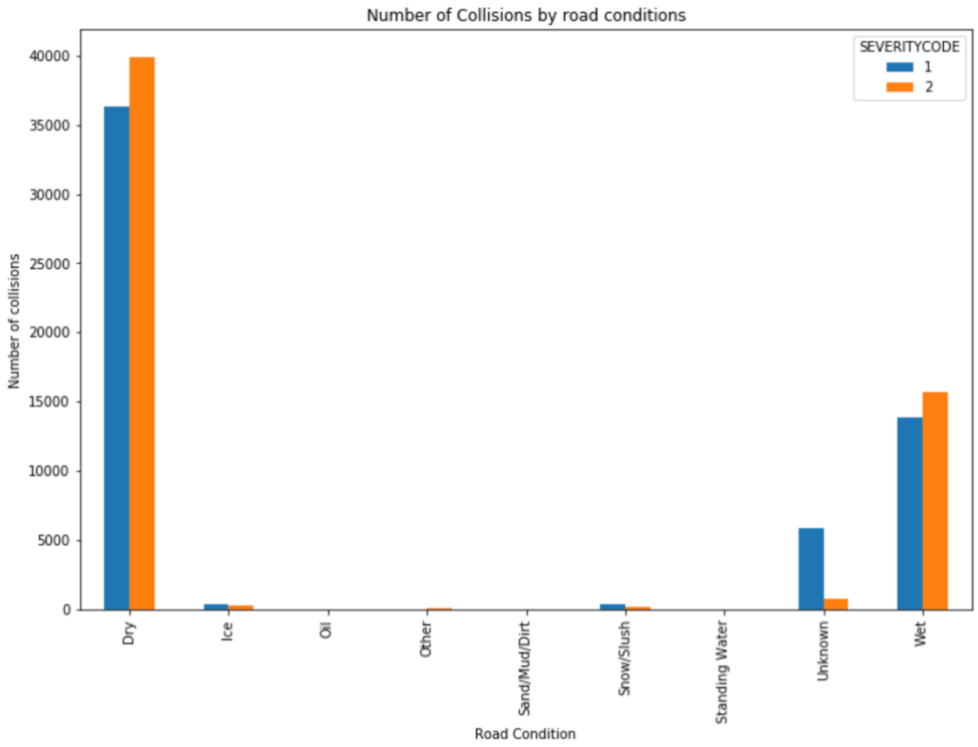

Figure 6: Collisions by Road Condition

3.1.5. Collisions by Light Condition

Data shows that most collisions happen during daylight or when it's dark and there are streets lights on. In daylight is more common to see type 2 severity code collisions, meaning, collisions resulting in injuries. But when it's dark and there are streets lights on, the severity of the collisions is almost equally common.

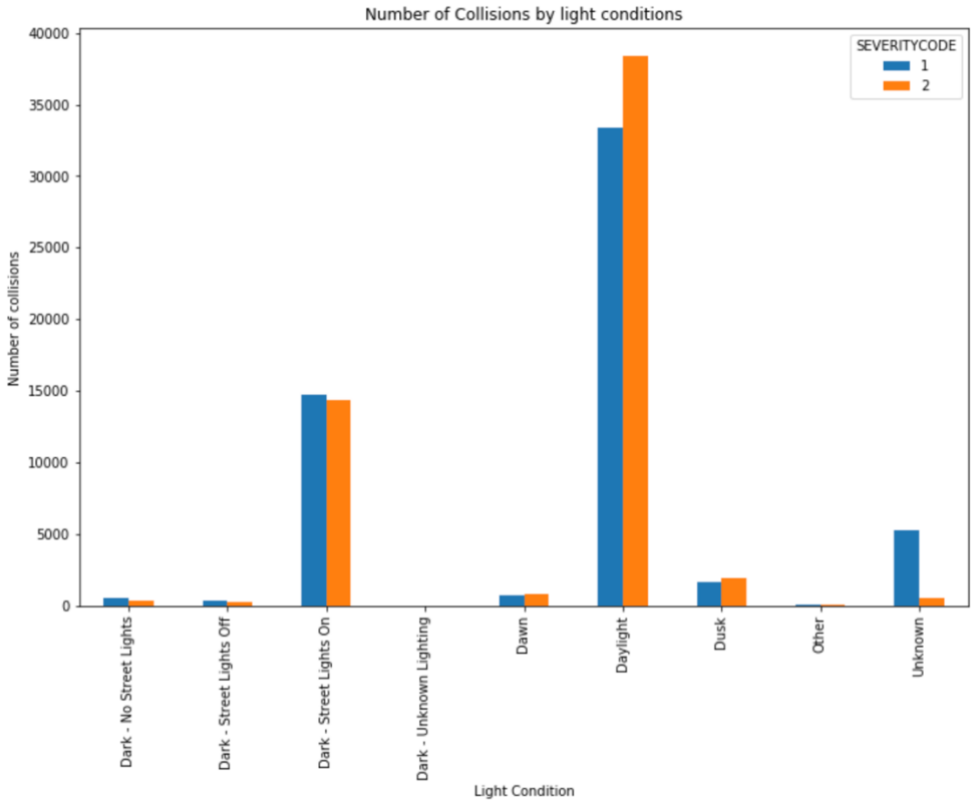

Figure 7: Collisions by Light Condition

3.1.6. Collisions by People Involved

Data shows that most collisions involve less than 7 people, but most common collisions involve 2 people. We can appreciate that when collisions involve two or less people, is more common that said collisions result only in property damage. However, when collisions involve between 7 and 10 people, is more common to see injuries.

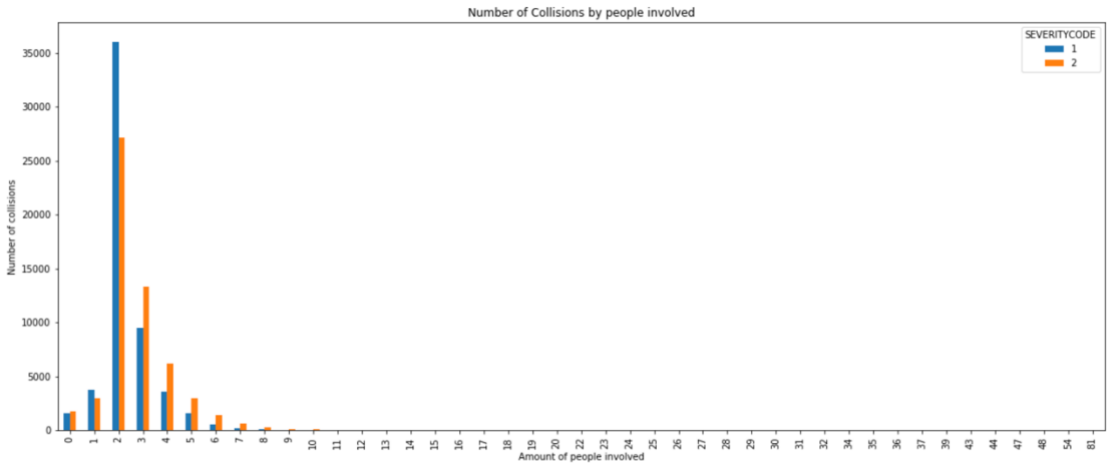

Figure 8: Collisions by People Involved

3.1.7. Collisions by Vehicles Involved

Data shows that most collisions involve less than 5 vehicles. However, in most collisions there are 2 vehicles involved with only property damage. However if the number of vehicles is 1,3,4, or 5, it is more common for the collision to result in injuries.

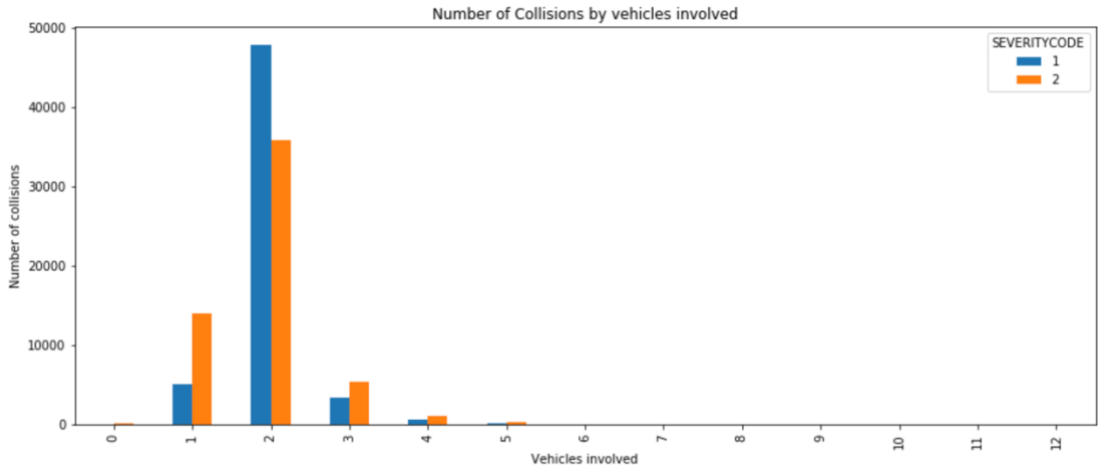

Figure 9: Collisions by Vehicles Involved

3.2. Modeling

Classification algorithms were considered to predict the severity of a collision. The dataset was split into a training and testing samples. Training sample consisted in 80% of the original dataset and testing sample consisted in the remaining 20% of the data. Four machine learning classification algorithms were created: K-Nearest Neighbors, Decision Tree, Logistic Regression, and Support Vector Machine.

3.2.1. K-Nearest Neighbors (KNN)

Due to size of the dataset it was not feasible to try several numbers of clusters and compare their accuracy, since this was a too time-consuming task. To solve this, the algorithm was computed with 2 clusters. This number was selected because it is the number of labels in the target data.

After running the algorithm, train set accuracy was 64.8% and test set accuracy was 62.4%.

3.2.2. Decision Tree

The *max_depth* parameter in the decision tree algorithm was optimized, considering a maximum of 50. The algorithm has optimized accuracy when *max_depth* was 11.

After running the algorithm, train set accuracy was 71.6% and test set accuracy was 71.2%.

3.2.3. Logistic Regression

The resulting model has a 75% precision for detection of label 1, property damage, and a 67% precision for detection of label 2, injuries.

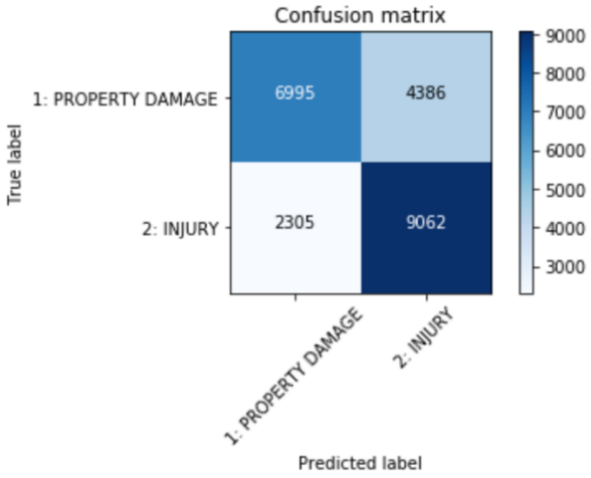

Figure 10: Confusion Matrix for Logistic Regression

3.2.4. Support Vector Machine (SVM)

The resulting model has a 74% precision for detection of label 1, property damage, and a 68% precision for detection of label 2, injuries.

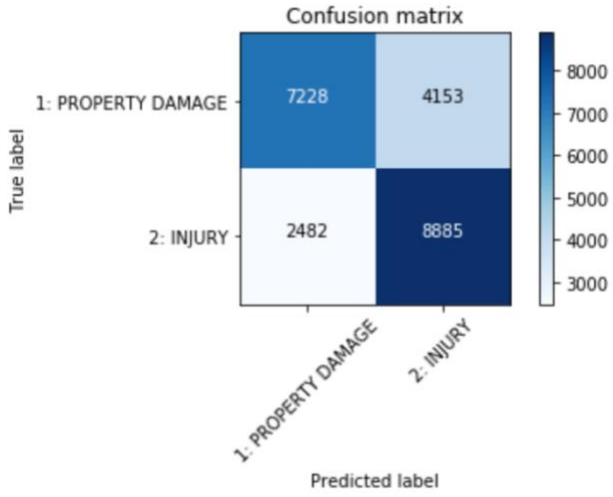

Figure 11: Confusion Matrix for SVM.

4. Results

The following Table shows the performance of the implemented algorithms.

ALGORITHM	F1-SCORE	JACCARD	LOG-LOSS
KNN	0.61	0.62	NA
DECISION TREE	0.71	0.71	NA
LOGISTIC REGRESSION	0.70	0.71	0.54
SVM	0.71	0.71	NA

Table 2: Results

5. Discussion

The overall goal of creating a model that could predict the severity of a car collision given general conditions based on historical data was achieved. The models created have a relatively good performance. However, it is difficult to choose one model in particular to use to predict collision severity since the

Decision Tree, Logistic Regression, and Support Vector Machine have almost identical performances.

K-Nearest Neighbors lower performance could be caused by the number of clusters chosen. This performance could be improved by tuning the parameter, which wasn't possible due to the processing power available and time constraints.

All three of the best-performing algorithms could be used by the Seattle Government to predict car collision severity with equally good results.

6. Conclusion

In this study, the goal was to predict the severity of a car collision given the general conditions during the time of the collision. The fields determined most important for the goal, and thus, the ones used, were the severity of the collision, number of people involved, number of vehicles involved, address type, collision type, weather condition, road condition, and light conditions.

Classification models were used to predict the severity of a collision, resulting in similar performances. These algorithms could be very useful in helping Seattle Government to create safety legislations or safety campaigns.