Report (Car Accident severity)

Introduction

Background

Seattle is the largest city in the state of Washington that has around 745000 people. The rate of people growth is 2.1% where is the second-fastest growth in US cities. Also, the number of tourism people is increased. Also, the number of workers is increased. So, all these factors could affect transportation where the number of cars is increased which lead to an increase in the number of accidents. Some of the accidents are dangerous and some not. It is important to know the severity of accidents and what are factors that affect them by predicting the severity of car accidents in future by using data of previous accidents.

Problem

Data that might contribute to determining the severity of accidents might include weather, speed, road condition, the light of road, address type, and the pedestrian right of way was granted or not. The aim of the project is to predict whether of car severity of accidents.

Interest

Seattle department of transportation would be very interested in accurate prediction of SEVERITY of accidents of cars. Also, this prediction could help to solve some factors that affect car accidents.

Data acquisition and cleaning

Data sources

Transportation SeattleCityGIS has shared the dataset in its website. Dataset contents data of report number, location, rain, weather, date, severity code, road condition, light condition, speeding, State Collision Code and so on.

Data Cleaning

There were a lot of missing values that could affect the prediction. I decided to add the results of some missing values of data that can add them or change. For example, the column of SPEEDING has yes, or No value and some records have a null value. So, I change the null value to No. This process does it with INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, and JUNCTIONTYPE columns.

Another clean data method is removed all rows that has null values and cannot edit the value of columns such as the column of State Collision Code.

Also, I have used column of date to get the year of the accident as well as the day of the accidents are weekends or weekdays.

After that, I will change the columns name to be clear and I can use them for exploring data.

Feature Selection

When I cleaned data, I got 187504 sample and 38 features. There are some redundant features that could not use for predication. For example, duplicate columns such as severity code and date timestamp. Also, I dropped some columns that could not use as features such as report

number, longitude, latitude, unique key for the incident, Secondary key for the incident, Key that corresponds to the intersection associated with a collision, and so on. After this process, I got around 19 columns. It could be changed when I start experiment.

Exploratory Data Analysis

Calculation of target variable

Number of severity accidents by year was not a feature in the dataset and had to be calculated. I choose to calculate number of accidents in each year from 2004 to 2020 to see how we can reduce this number of accident (see FIGURE 1). The factors affect severity of accidents are speeding, weekends or weekdays, inattention, alcohol or drags abuse, parked cars, weather, address types, road condition, light condition, junction type, and collision type.

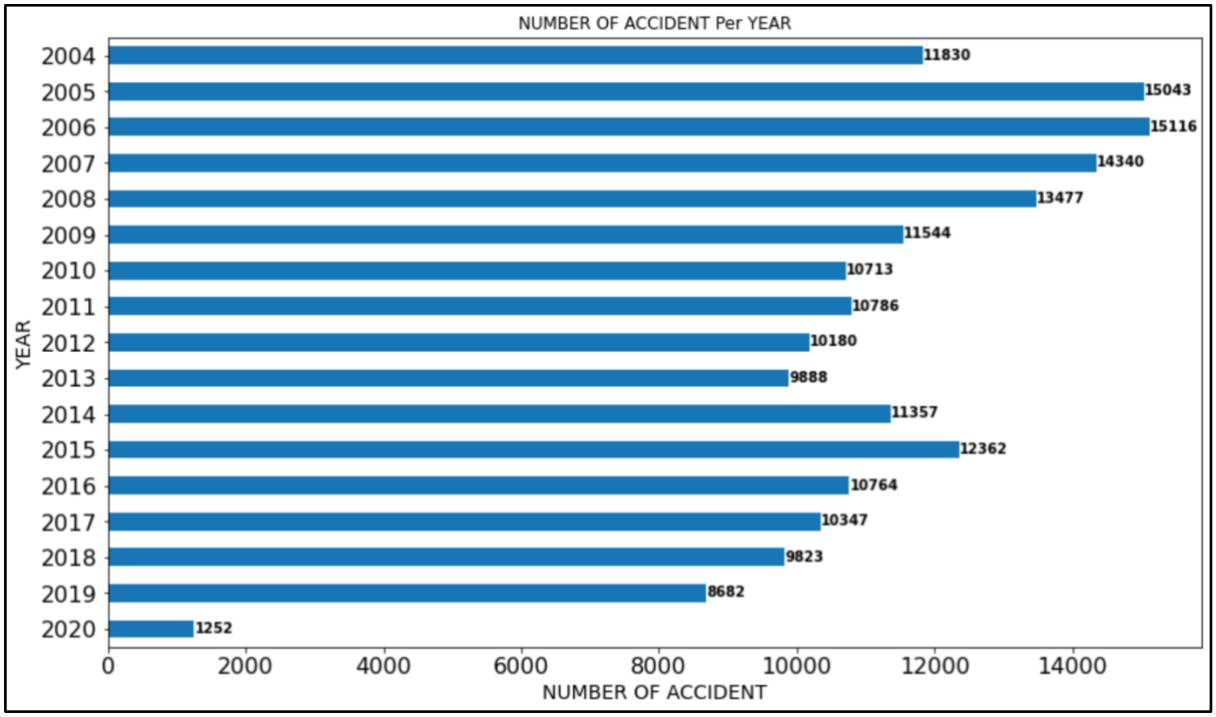

FIGURE 1: Number Of Accidents from 2004 and 2020

Also, it can be seen the number of each type of severity (see Figure 2). The number of injury collision is 56870 accidents. However, the number of property damage only collision is 130634 accidents.

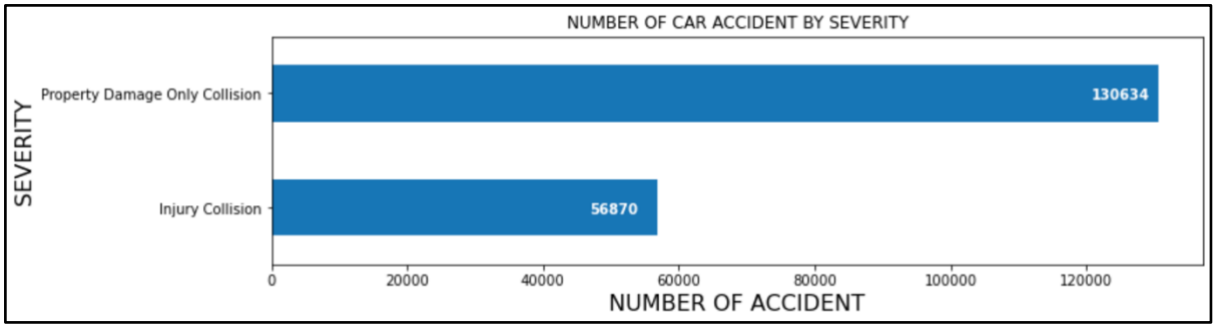

FIGURE 2: Number of Accidents by SEVERITY

Relationship between severity of accidents and Weekdays and Weekend

The number of severity of accidents was increased in weekdays. If you see figure 3, the number of accidents in weekend is 47815 accidents and the number of accidents in weekdays is 139689 accidents. The justification for this number is due to the working hours in weekdays.

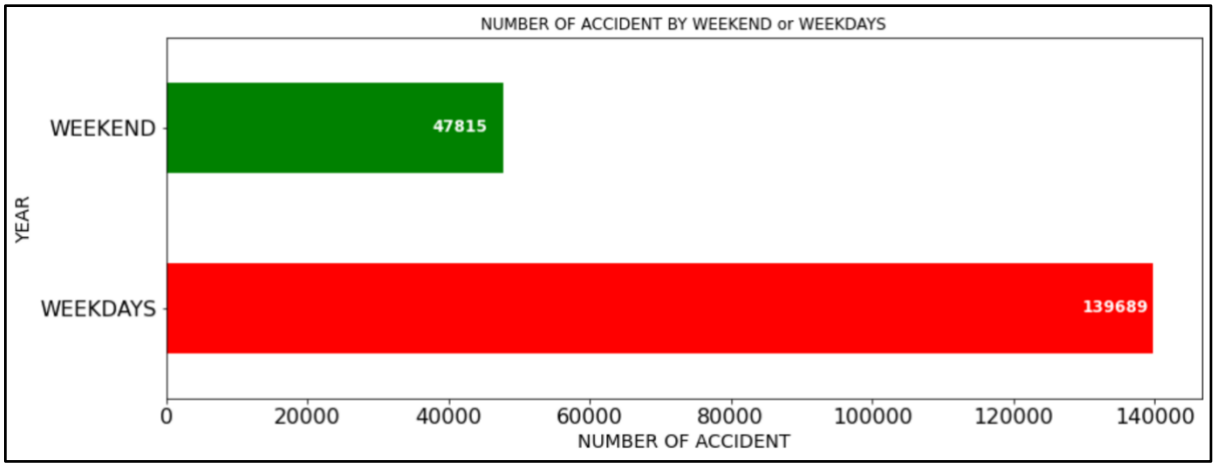

FIGURE 3: The Number of accidents by Weekend or Weekdays

Also, it can be seen figure 4 The number of injury collision in weekend is less than the number of accidents injury collision in weekdays which are 13847 and 43023 accidents respectively. However, the number of property damage only collision weekdays is highest which is 96666 accidents.

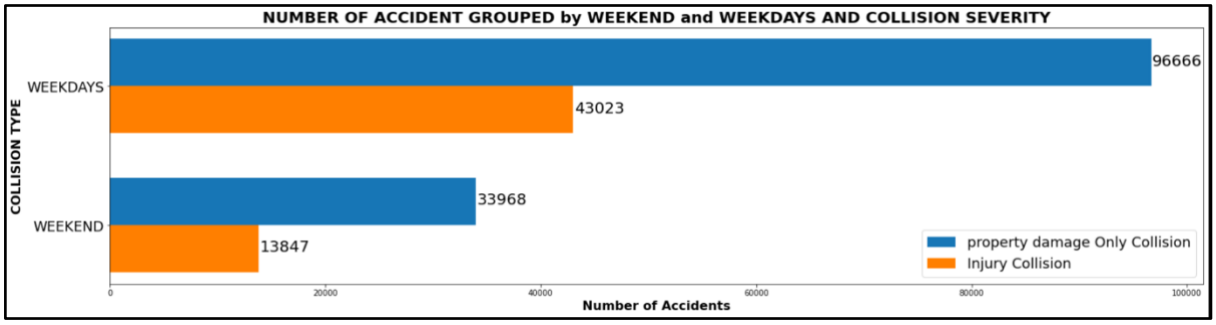

FIGURE 4: Number of severity accidents in weekdays and weekend

Relationship between severity of accidents and Weather

Weather could affect the severity of accidents. In figure 5, the highest category of weather is clear which has the highest number of accidents.

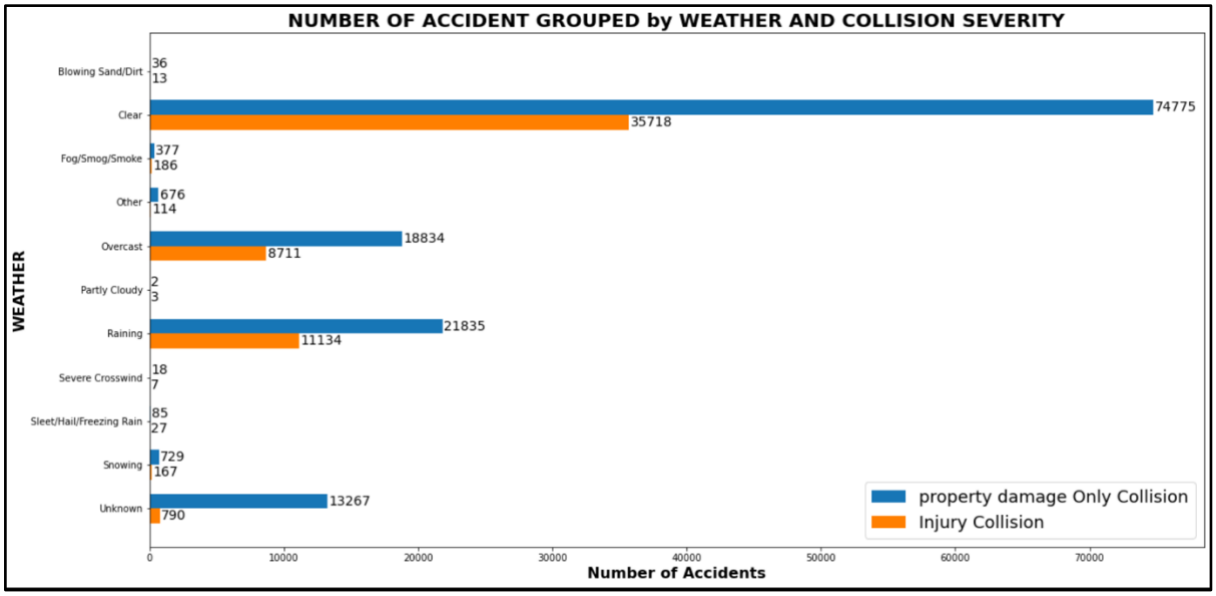

FIGURE 5: Number of accidents grouped by weather and collision severity

Relationship between severity of accidents and road condition

Condition of road such as Dry, Ice, Oil, Sand/Mud/Dirt, Snow/Slush, Standing Water, Wet could affect on accident severity. In figure 6, it can be seen the number of each type of severity in each condition of road. It appears that dry and wet condition have the greatest number of accidents.

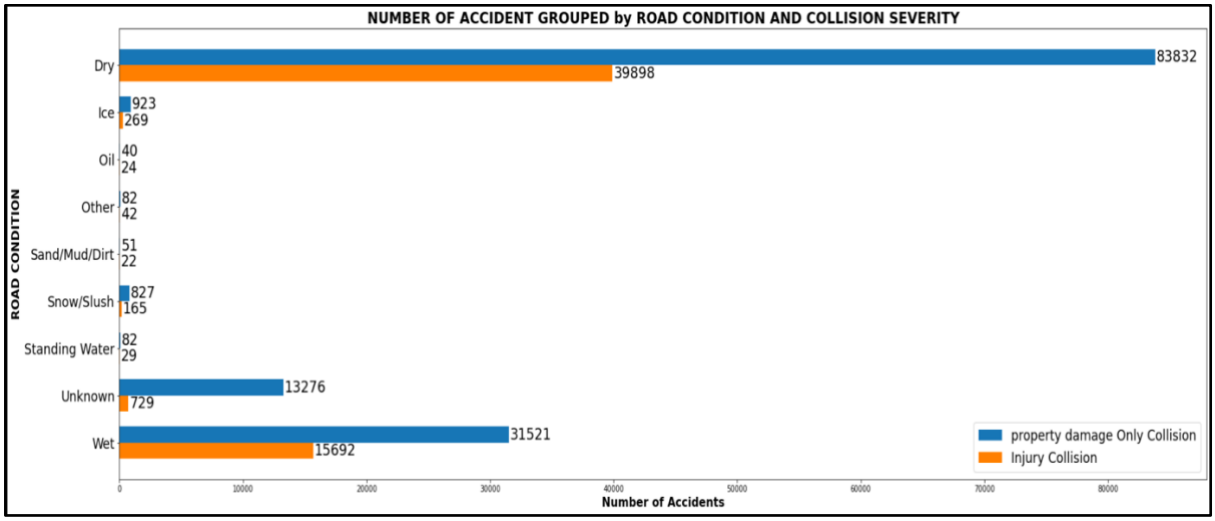

FIGURE 6: number of accidents by condition of road

Relationship between severity of accidents and light condition

Light condition affects the accidents number. Figure 7 displays the number of each type of severity accidents in each category of light condition such as dark with light street on, dark with light street off, daylight, and so on.

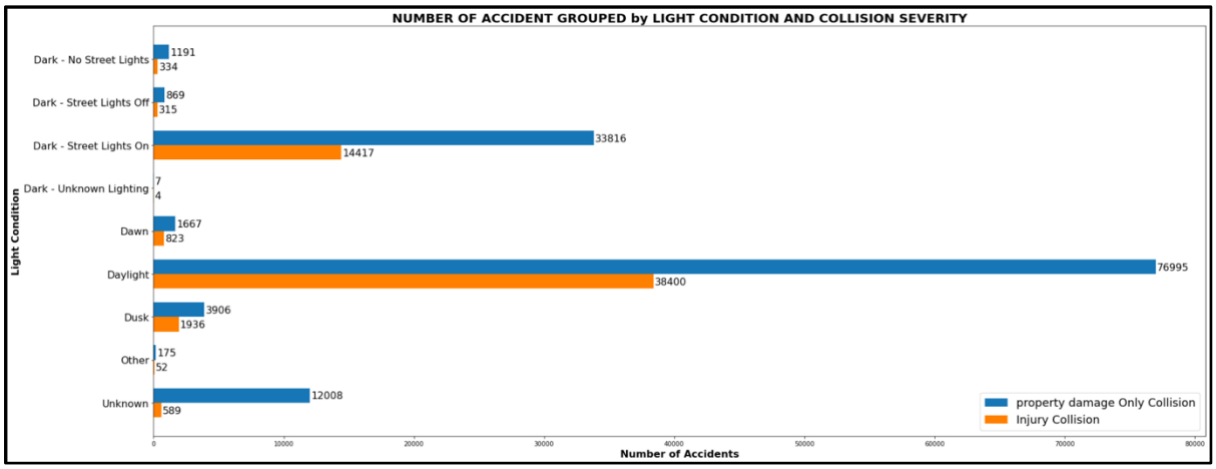

FIGURE 7: Number of accident with light condition

Relationship between severity of accidents and collision type

Type of collision illustrate behaviours of drivers. Figure 8 shows the number of types of severity collision that grouped by type of collision. The highest type of collision in property damage only collision is parked car which is 44031 collision. However, the highest type of injury collision is rear ended which has 14565.

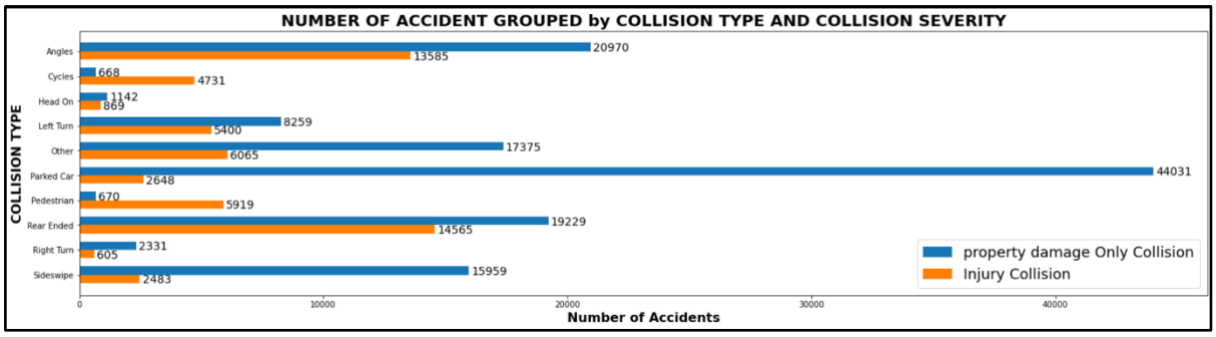

FIGURE 8: Number of Accidents by collision type

Relationship between severity of accidents and speeding

People think that the speed factor may have the highest impact on collision severity. However, figure 9 illustrates number of accidents severity with speeding or not. As it can be seen in figure 9, the most accidents are not affected by speeding. Also, most of accidents that affected by speeding is property damage only.

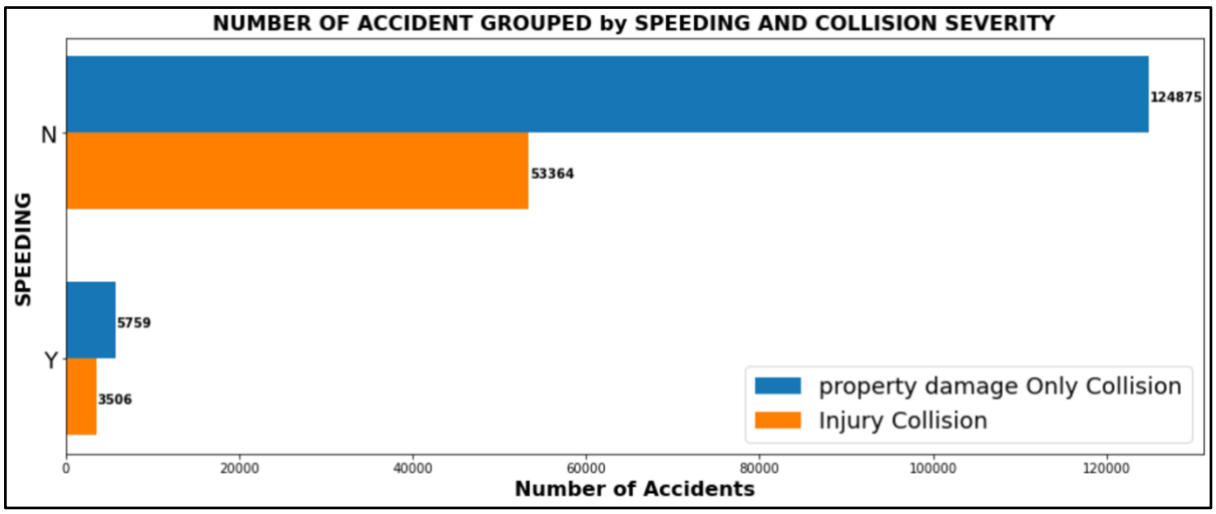

FIGURE 9: Number of Accidents by speeding

Predictive Modelling

The dataset of accidents severity has two classes which are property damage only, and injury collision. The model that can use with this dataset is classification model such as K nearest neighbour, decision tree, SVM, logistic regression, and naïve bayes.

Classification Models

The application of classification models is clearer. The dataset has imbalanced classes (See Figure 2). So, I chose to divide the dataset into two sets, one for training which is 70% and others to test which is 30%. Also, I chose F1-Score as the metric here because the results would probably be presented with binary and F1-Score puts more emphasis on the precision and recall together than other metrics. Decision Tree, Logistic Regression, SVM, Random Forest, K Nearest Neighbour models and Naive Bayes model were tuned and built. The best performance of these models is 7-Nearest neighbour where is 0.7116 (see Table 1). However, the best accuracy is for SVM where is 0.7479 (see table 1). So, figure 9 shows the confusion matrix of 7-Nearest Neighbour and it can be seen how many true positive, false positive, true negative, and false negative.

	Logistic Regression	Decision Tree	Random Forest	Naïve Bayes	7-Nearest Neighbour	SVM
F1-Score	0.6916	0.6840	0.6900	0.6317	**0.7116**	0.6874
Accuracy	0.7474	0.7476	0.7446	0.6207	0.7148	**0.7479**
No. of True Positives	38366	**38835**	38177	20188	31828	38664
No. of False Positives	13351	13807	13319	**2302**	8646	13623
No. of False Negatives	855	**386**	1044	19033	7394	557
No. of True Negatives	3680	3224	3712	**14729**	8385	3408

Table 1: Performance of classification models. Best performance labelled in red.

I also chose to use the confusion matrix to evaluate the models. so, I chose to display the confusion matrix for the highest accuracy and the highest F1-score models. Figure 10 shows the confusion matrix of 7-Nearest Neighbour which has highest F1-score and Figure 11 shows the confusion matrix of SVM which has highest accuracy.

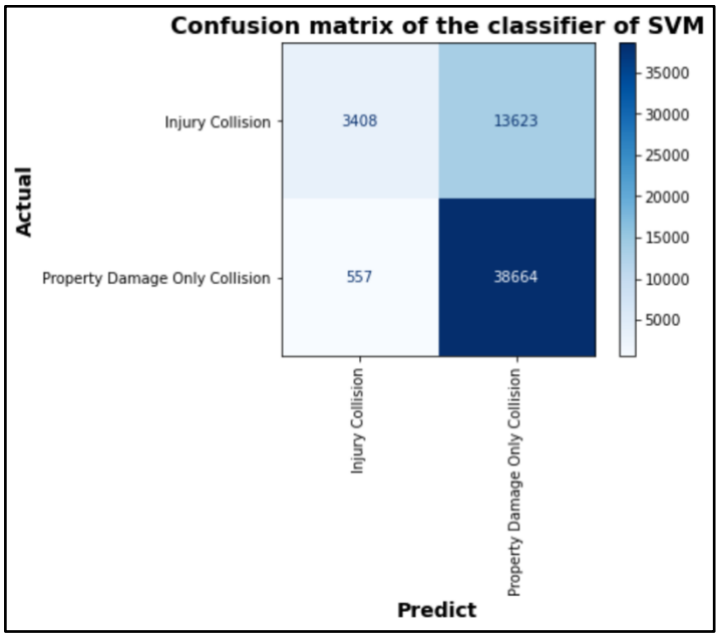

FIGURE 10: Confusion Matrix of SVM

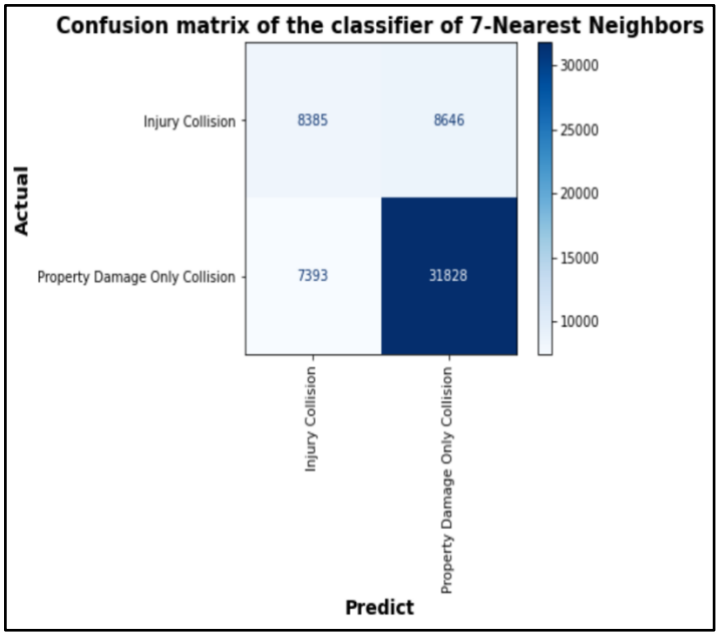

FIGURE 11: Confusion Matrix of 7-Nearest Neighbour

Conclusion and future

In conclusion, I analysed the relationship between accident severity and their factors could affect severity of accident such as weekdays or not, weather, road condition, light condition, collision type, and speeding. Then, I built the classification models that can be useful to predict the severity of accident. These could help Seattle department of transportation to control the factors that affect injury collision. To improve accuracy and F1-score, we need to add some more data to build balanced dataset.