Car Accident Report

MSc Mario Paredes Uriona

September, 2020

1 Introduction

1.1 Background

Car accident is one of the most common accident in the world, especially in crowded cities. There are different factors that are involved in this kind of accident, such as weather conditions, road conditions light conditions and more. The mix of these factors can obtain a higher probability to get an accident. Thus, it is important to analysis all the conditions to calculate this probability.

1.2 Problem

The number of car accident in Seattle is noticeable. Thus it is important to determine the probability to get an accident depending of the different conditions of the road, day and more.

2. Data Sources and Cleaning

2.1 Data Sources

Data acquired from the historical data from Seattle. this data base includes different attributes, such as location, date of accident, severity of the accident and more. The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present. The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

3. Exploratory Data Analysis

The are 37 attributes of which after an analysis, the Weather condition, light condition and road condition which were defined as independent variables. Then, the Severity code was defined as the dependent variable.

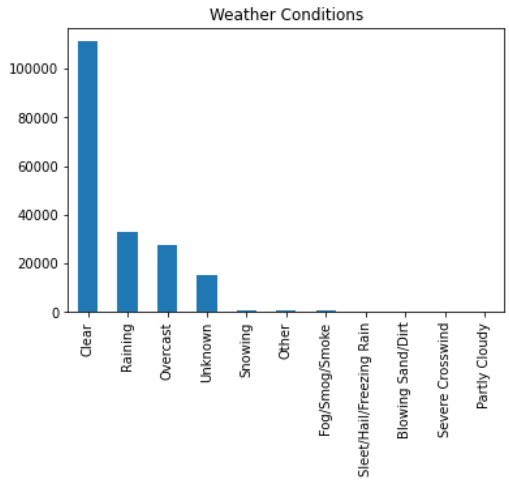

Figure 1.- Weather Condition

As it can be seen in Figure 1, Clear condition is the main attribute, which has the majority of the data. Most of the car accident occurred in clear light condition.

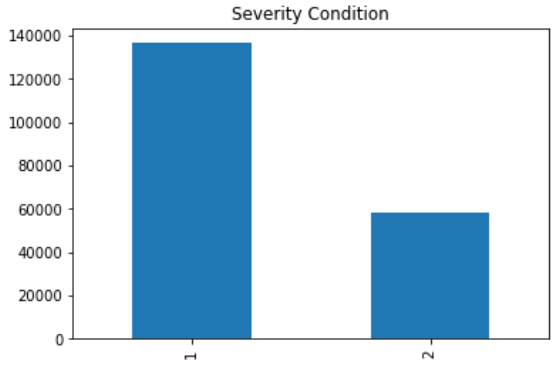

Figure 2.- Severity Condition

As shown in figure 2, almost 66% of the car accident registered are severe. For this reason, it was necessary to perform a data balancing. As there is a reasonable amount of data for the *Severity Condition 2*, the Under balance method was selected. As a result, the data for *Severity Condition 1* was randomly reduced to reach the same quantity data as the *Severity Condition 2.*

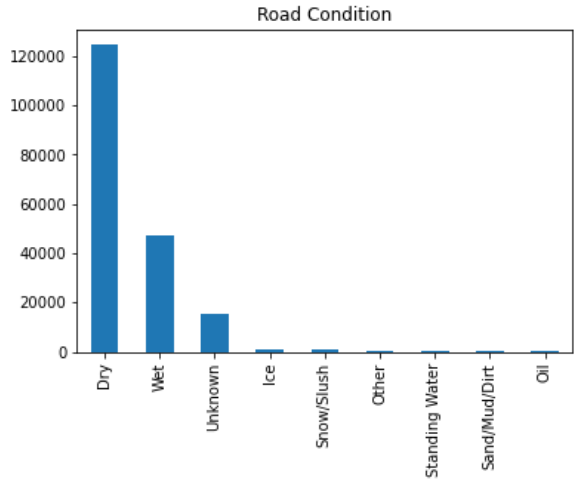

Figure 3.- Road Condition

In figure 3 shows the road condition of the accident. Dry road condition has the majority, followed by wet.

It was necessary to convert categorical data into numerical data, in order to perform the algorithm. The figure below displays the result of the conversion.

	ROADCOND	WEATHER	LIGHTCOND
0	Wet	Overcast	Dark - Street Lights On
1	Dry	Clear	Daylight
2	Dry	Clear	Dark - Street Lights On
3	Dry	Clear	Dark - Street Lights On
4	Dry	Clear	Daylight

	ROADCOND	WEATHER	LIGHTCOND
0	8	4	5
1	8	6	2
2	0	4	5
3	0	1	5
4	8	6	5

Figure 4.- Conversion categorical data into numerical data

Predictive Modelling

As or target data is the servility of the accident, the type of model is Classification.

K-Nearest Neighbors

The k-nearest neighbors (KNN) algorithm from supervised machine learning algorithm that can be used to solve both classification and regression problems.

K in KNN, is the number of nearest neighbors to be examined. To solve this algorithm, a part of the data is reserve for testing the accuracy of the model.

The other part of set to the training model, in which k =1 is used to train the model, and calculate the accuracy of prediction using all samples in your test set. The process is repeated by increasing the number of k, and finally the value of k is the one that best represent the accuracy of the model. Decision Tree It is a machine learning algorithm in which the algorithm is a decision tree. Logistic regression It

is a linear algorithm similar to linear regression, the difference is that the target value if a categorical value, instead of a numerical value in the linear regression.

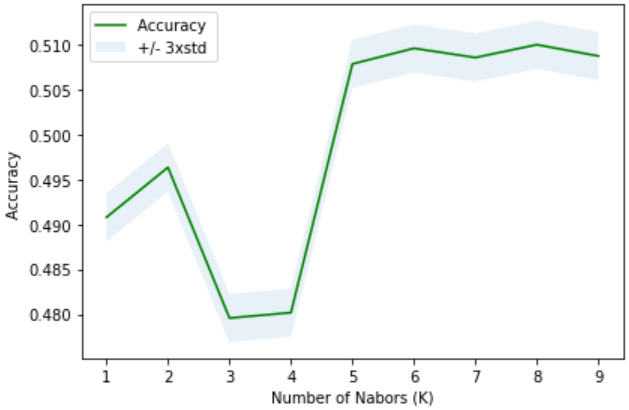

Figure 5.- KNN Plot Result

To summarize, the results of the KNN Classification model is shown is Table 1.

Table 1. KNN Results

Train Accuracy	0.50916
Test Accuracy	0.50960
F-1 Score	0.49953
Jaccard-Score	0.27193

Decision Tree

Decision Tree It is a machine learning algorithm in which the algorithm is a decision tree.

The results of the Decision Tree model is shown is Table 1.

Table 2. Decision Tree Model Results

Decision Tree Accuracy	0.52533
F-1 Score	0.49260
Jaccard-Score	0.26872

Logistic regression

Logistic regression It is a linear algorithm similar to linear regression, the difference is that the target value if a categorical value, instead of a numerical value in the linear regression.

Table 3. Logistic Regression Model Results

Decision Tree Accuracy	0.52198
F-1 Score	0.51599
Jaccard-Score	0.40015

Results

As it can be seen in figure, all the values of are similar for all the methods applied in this work. Although the accuracy and the F-1 Score results of the three models (KN, Decision Tree and Logistic Regression) are roughly similar, the Jaccard score for the Logistic Regression model is higher than the other models.

Table 4. Results

	KNN	Decision Tree	Logistic regression
Model Accuracy	0.50960	0.52533	0.52198
F-1 Score	0.49953	0.49260	0.51599
Jaccard-Score	0.27193	0.26872	0.40015

Conclusions

In this works, different classification methods were implemented into a set of car accident data from Seattle, in order to obtain the probability to suffer a car accident depending of different factor such as weather condition, road condition and light conditions. these parameters were defined as the main parameter for this study.

As a result, the algorithm to measure the probability to suffer an accident was developed with a considerable good accuracy.