Predicting the Car Accident Severity in Seattle

Sakif Ahmed

September 21, 2020

1. Introduction

Traffic accidents has become a major headache for the governments around the world. A global status report on traffic safety notes that there were 1.25 million traffic deaths in 2013 alone, with deaths increasing in 68 countries when compared to 2010. Thus, the importance for accident prediction is felt in various fields: optimizing public transportation, enabling safer routes, and cost-effectively improving the transportation infrastructure; all in order to make the roads safer. Accident analysis and prediction has been a topic of much research in the past few decades. Analyzing the impact of environmental stimuli (e.g., road-network properties, weather, and traffic) on traffic accident occurrence patterns, predicting frequency of accidents within a geographical region, and predicting risk of accidents are the major related research categories.

The objective of this project is to use the dataset on car accidents on Seattle, USA to explore the incident severity and to build a machine learning algorithm to predict the future accident severity so that it can limit the future possible accidents. Python is used for this project to analyze the dataset and to build a fitting model for the objective.

2. Data sources and Data Preprocessing

2.1 Data Sources

The example dataset comes from the web page of Coursera Course Applied Data Science Capstone. The .csv file contains 19,4673 pieces of collision records of Seattle. The data for this capstone project is offered by the SDOT Traffic Management Division and recorded by Traffic Records Group. It covers the annual collisions data from 2004 to the present. The time-frequency

of this dataset is weekly and it shows the traffic collision records in Seattle. The example datasets contain 194,673 pieces of records starting from 2004. The attributes in the datasets cover the weather condition, road condition, collision type and fatality.

The other relevant data comes from the database of Seattle government. The main data source is from the Seattle Department of Transportation. The basic background information is offered by this department.

2.2 Data Cleaning

The original dataset is not suitable to work with. There are 37 attributes with a lot of missing values. To clean the dataset, the rows with missing values were dropped. After cleaning the dataset from missing values, the dataset now contains 110,586 pieces of collision records.

2.3 Feature Selection

After data cleaning, there were 110,586 samples and 37 features in the data. Upon examining the meaning of each feature, I have decided to focus on only six features: SEVERITYCODE, JUNCTIONTYPE, PERSONCOUNT, WEATHER, ROADCOND, LIGHTCOND. These features were selected due to their suitability for the upcoming building of the machine learning models. In addition, I have selected SEVERITYCODE as the target variable as it is best suited for car accident severity prediction.

2.3 Feature Engineering and Balancing of Data

The selected features have two types of variables: numerical and categorical. To best suit the machine learning model, I have converted the categorical features to numerical values. This action was done by using the df.replace() method.

On the other hand, the targe feature SEVERITYCODE has a problem too. The data set of this feature is imbalanced (Figure.1(a)). This imbalance may lead the machine learning model to be biased. This problem can be fixed by down sampling the majority class by using the resample method (Figure.1(b)).

Table 1. Description of the selected features

Feature	Data Type, Length	Description
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury • 2—injury • 1—prop damage • 0—unknown
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
PERSONCOUNT	Double	The total number of people involved in the collision
WEATHER	Text, 300	A description of the weather conditions during the time of the collision
ROADCOND	Text, 300	The condition of the road during the collision
LIGHTCOND	Text, 300	The light conditions during the collision

SEVERITYCODE	
1	136485
2	58188

(a)

SEVERITYCODE	
2	58188
1	58188

(b)

Figure. 1 The values of feature ser SEVERITYCODE: (a) imbalanced (b) after down sampling

3. Exploratory Data Analysis

To explore the relationship among the feature sets, I have conducted exploratory data analysis ploting the feature variable against the target variable. To achive effective visualization of the features, bar plot is used.

3.1 Relationship between Junction type and Severity

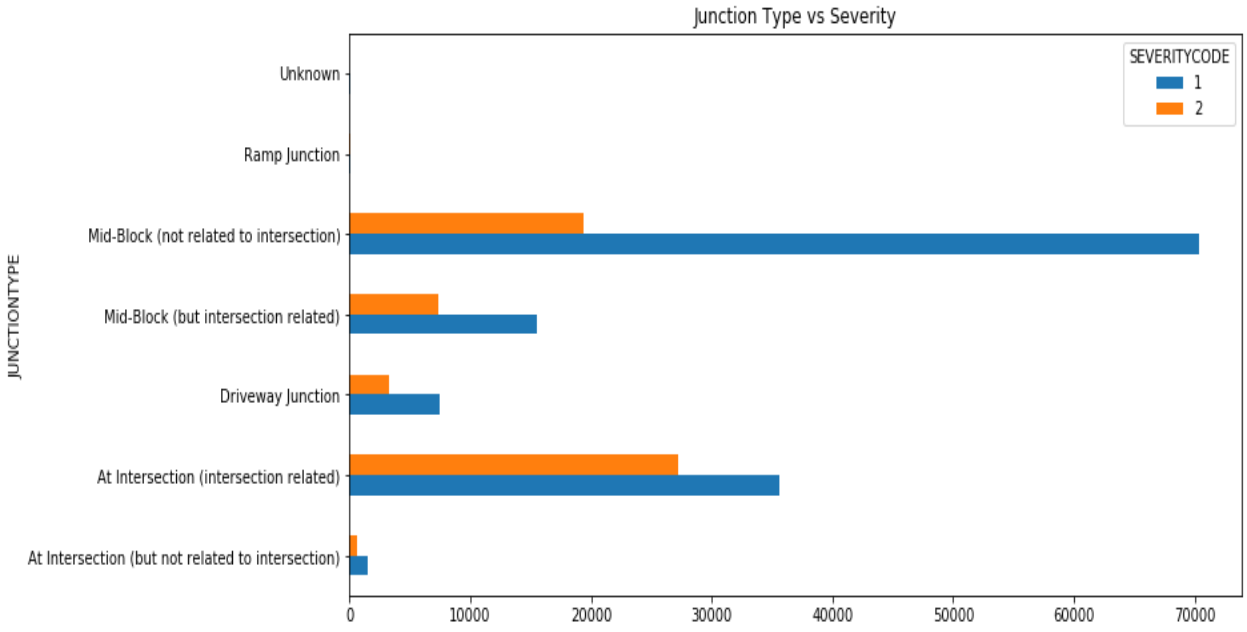

Figure.2 Visual representation of the relationship between junction type and severity

The feature junction type is a categorical variable which refers to the category of junction at which collision took place. It is obvious that there are some junctions more prone to car collision than other, most notably Mid-Block (not related to intersection) and At Intersection (intersection related) (Figure.2). Furthermore, we can also find from the visual representation of the relationship that Mid-Block (not related to intersection) junction has very high property damage. On the other hand, although At Intersection (intersection related) has lower property damage, this junction has the highest number of accident injuries.

3.2 Relationship between Number of Persons and Severity

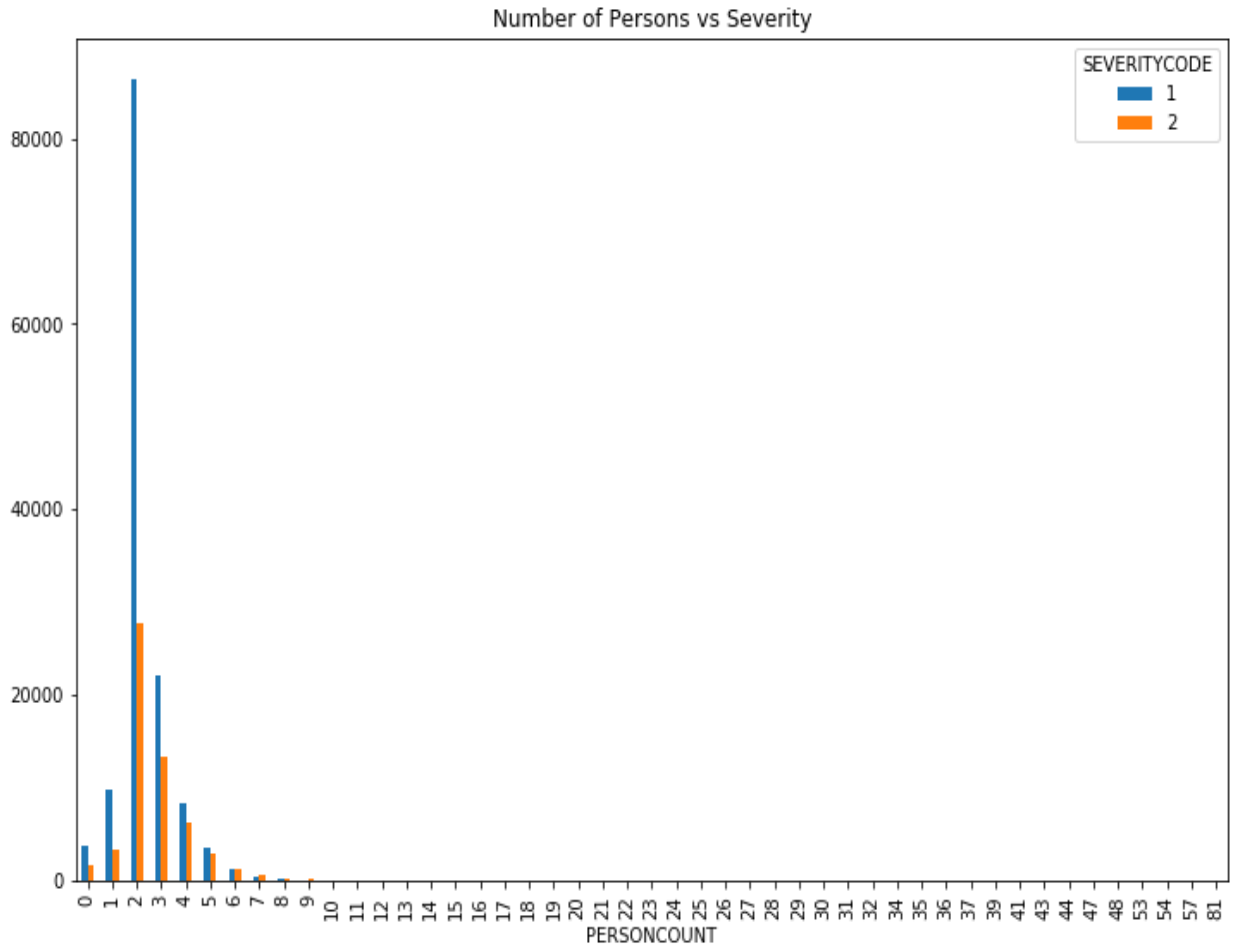

Figure.3 Visual representation of the relationship between number of persons and severity

The feature which denotes the total number of people involved in the collision in the dataset is known as PERSONCOUNT. From the visual representation of the relationship between number of persons and severity (Figure.3), it can be stated that car accidents involving 2 persons is quite severe, both on property damage and injuries. This visualization can also be an indicator to avoid driving alone or with a second person in the city as the maximum amount of car accident involves 2 persons. This indicator can be enhanced if there was an attribute whether the driver alone or with someone.

3.3 Relationship between Weather and Severity

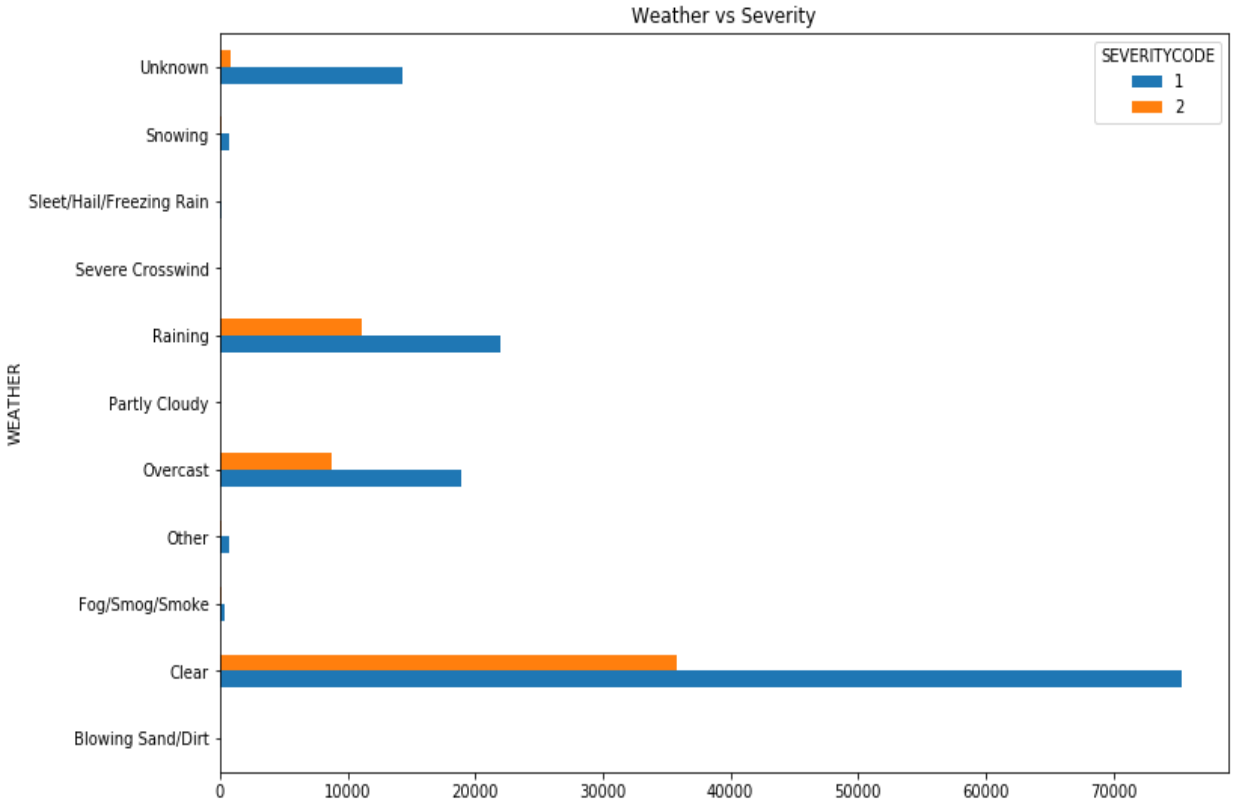

Figure.4 Visual representation of the relationship between Weather and severity

Weather is one of the most powerful attributes in this dataset to compare with car accident severity. It is categorical variable that provides a description of the weather conditions during the time of the collision. The visualization of this relationship of weather and severity provided rather an odd occurrence (Figure.4). From the visualization, it is quite clear that rather than in bad weather conditions, most car accidents occur in clear weather conditions. Both, the property damage and the incident of injury is highest in the clear weather conditions. This may be due to the low traffic movement during the bad weather conditions.

3.4 Relationship between Light Conditions and Severity

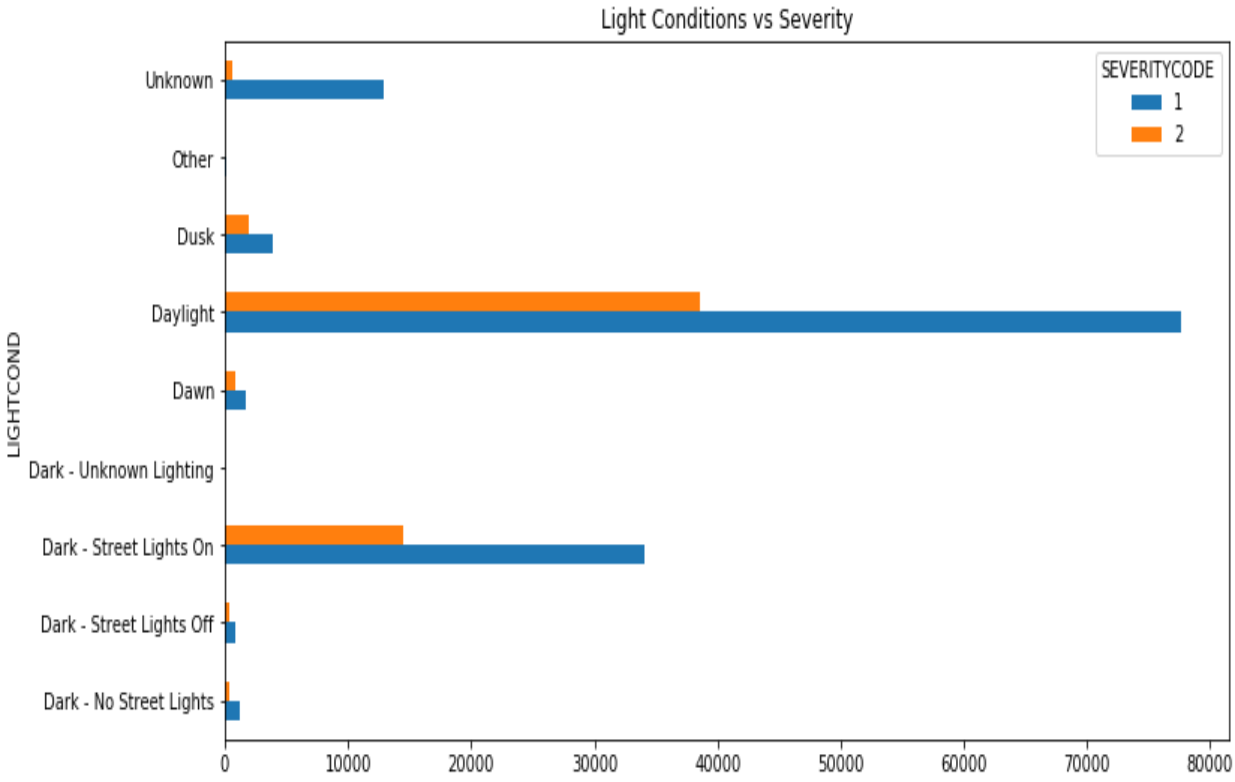

Figure.5 Visual representation of the relationship between Light conditions and severity

Light condition is a categorical feature that describes light conditions during the collision. This feature is quite important to compare with car accident severity. From Figure.5 the visual representation shows that most car accidents occur during the daylight. In addition, it can be also stated that property damage is quite high in daylight car accidents rather than in darker light conditions.

3.5 Relationship between Road Conditions and Severity

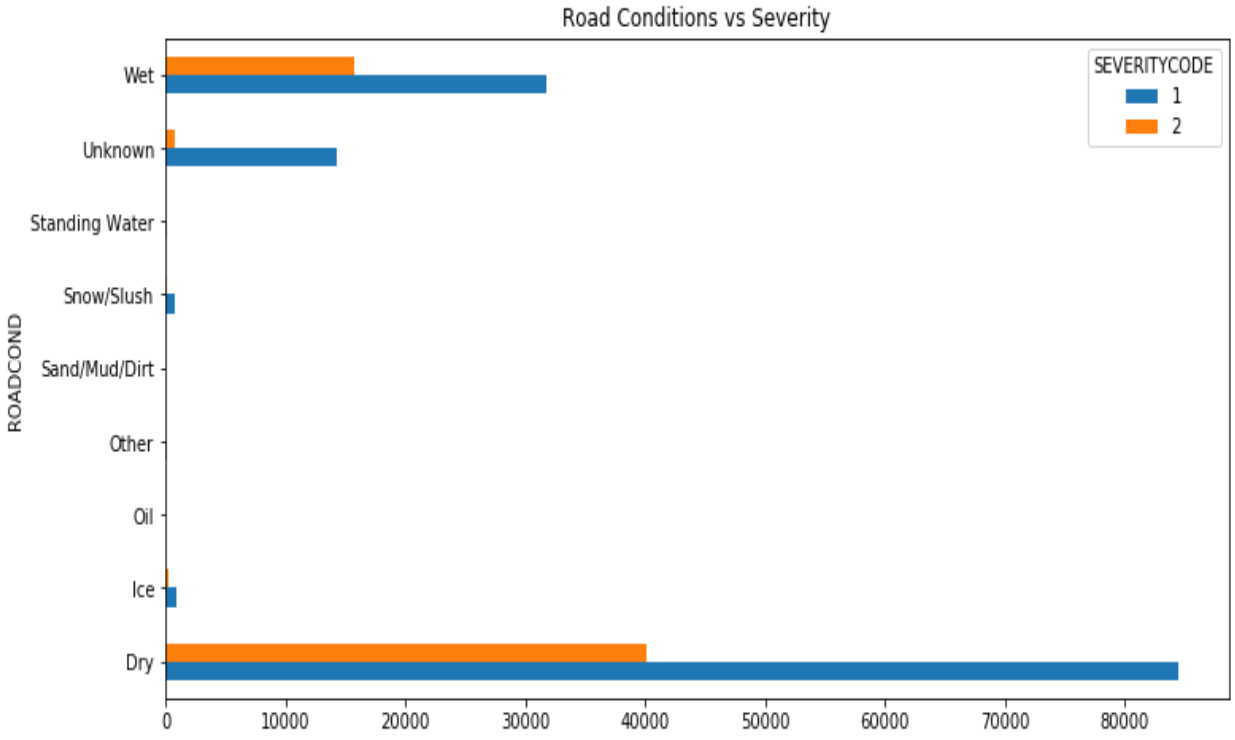

Figure.6 Visual representation of the relationship between Road conditions and severity

Road condition is another attribute that can be effectively plot with the target feature severity. This variable describes the condition of the road during the collision. From the visualization in Figure 6, it can be stated that car accidents happen in mainly two road conditions: wet and dry. In both cases the property damage was high. It can also be denoted that, most car accidents occurred in dry road conditions.

4. Building the Machine learning Models

To prepare the final balanced dataset for the machine learning models, the features are divided into train and test set by using train_test_split from sklearn.model_selection. The following machine learning algorithms were used in this project:

1. K Nearest Neighbor (KNN)

2. Support Vector Machine (SVM)

3. Decision Trees

4. Logistic Regression

4.1 K Nearest Neighbor (KNN)

K-Nearest Neighbors is an algorithm for supervised learning. Where the data is trained with data points corresponding to their classification. Once a point is to be predicted, it takes into account the 'K' nearest points to it to determine the classification. In this sense, it is important to consider the value of k. It considers the 'K' Nearest Neighbors (points) when it predicts the classification of the test point.

4.2 Support Vector Machine (SVM)

SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data is transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.

4.3 Decision Trees

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

4.4 Logistic Regression

Logistic Regression is a variation of Linear Regression, useful when the observed dependent variable is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables.

Logistic regression fits a special s-shaped curve by taking the linear regression and transforming the numeric estimate into a probability with the following function, which is called sigmoid function.

5. Evaluation of the Machine Learning Algorithms

To evaluate the machine learning models, Jaccard index and F1 score were used. Jaccard index can be defined as the size of the intersection divided by the size of the union of two label sets. If the entire set of predicted labels for a sample strictly match with the true set of labels, then the subset accuracy is 1; otherwise it is 0.

F1 score is the harmonic average of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. It is a good way to show that a classifier has a good value for both recall and precision. Precision is a measure of the accuracy provided that a class label has been predicted. It is defined by:

$$\text{Precision} = TP / (TP + FP)$$

TP = True Positive

FP = False Negative

Recall is true positive rate. It is defined as:

$$\text{Recall} = TP / (TP + FN)$$

TP = True Positive

FP = False Negative

The calculated accuracy of the machine learning models is listed in Table.2. From the accuracy tests it was found that Support Vector Machine (SVM) and Decision Trees obtained the highest accuracy scores among the machine learning models. These two models scored 0.63 on both Jaccard index and F1 score.

Table.2 Jaccard index and F1 score of the machine learning models

Machine Learning Algorithm	Jaccard index	F1 score
K Nearest Neighbor (KNN)	0.59	0.57
Support Vector Machine (SVM)	0.63	0.63
Decision Trees	0.63	0.63
Logistic Regression	0.58	0.58

6. Conclusion

In conclusion, it can be stated that the machine learning models, Support Vector Machine (SVM) and Decision Trees are both suitable for predicting the car accident severity in Seattle. These models can be used in mobile apps to alert the user of daily car accident risks and thus providing more safety than before. However, this model can be improved in several ways. One interesting thing about this dataset is that most car accidents occurred in ideal driving conditions. If an attribute that describes about the specific reasons for the accidents in these ideal conditions can be delivered with this dataset, then the machine learning model can be even more effective in predicting the severity.