

**Accident Severity Detection:  
A Case Study in the United States**

By Ghita Benboubker

September 5<sup>th</sup>, 2020

## **Abstract**

Road accidents are still prevalent nowadays and their severity varies. If we can predict the severity of an accident before it happens, we may better estimate human injuries and material damages, help emergency responders provide faster and more appropriate care, and reduce traffic congestions. In the present paper, we take the United States as a case study in predicting the severity of accidents based on studied attributes. To that end, we extensively explore our dataset, and compare two classification models, namely: Decision Trees and Logistic Regression. In evaluating the performance of each machine learning models, we use the Jaccard Index, the F1-Score, and the Log Loss (only relevant for Logistic Regression). We find that, for our data set and for the preprocessing steps followed, the model with the best scores is Logistic Regression.

## Table of Contents

<b>Abstract .....</b>	<b>2</b>
<b>Introduction.....</b>	<b>4</b>
<b>Methodology .....</b>	<b>4</b>
<b>1. Description of the Data .....</b>	<b>4</b>
<b>2. Approach.....</b>	<b>4</b>
<b>3. Data Preprocessing .....</b>	<b>10</b>
<b>4. Modelling and Predictive Analysis .....</b>	<b>12</b>
Decision Trees Classification.....	13
Logistic Regression .....	13
<b>Results &amp; Discussion .....</b>	<b>13</b>
<b>Conclusion .....</b>	<b>13</b>
<b>References.....</b>	<b>13</b>

# Introduction

Road transportation, while convenient, gives rise to road accidents with varying severity. According to The Organization for Economic Co-operation and Development (OECD), in 2017, the United States registered 1.923 million road accidents involving casualties [1]. These accidents may happen for several reasons either pertaining to the driver, the vehicle(s), and/or weather conditions. While the most notable repercussions of road accidents are fatal and non-fatal injuries to the people involved, as well as the wrecking of the vehicle(s), traffic congestions also occur as a result. The extent of these three consequences depends on the severity of the accident. Therefore, predicting the severity of an accident may facilitate estimation of the human and material harm or losses, help emergency responders provide care effectively and efficiently, thus also reducing traffic congestions.

## Methodology

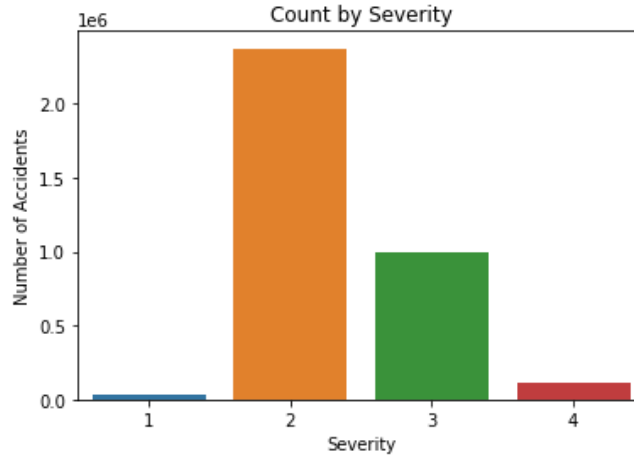
### 1. Description of the Data

The raw data set employed was downloaded from Kaggle, a data science website hosting open source data sets (<https://www.kaggle.com/sobhanmoosavi/us-accidents/download>). The data set contains 3.5 million records of countrywide road accidents in the United States, and 47 attributes (excluding the identification column and the road severity label column). The label column (i.e. Severity) has values ranging from 1-4, where 4 is the highest accident severity. The observations cover about 49 states and span a period from February 2016 to June 2020. The data was collected from sources such as traffic cameras, traffic sensors, law enforcement agencies, and the US Department of Transportation.

### 2. Approach

In our approach to analyze our dataset, we first start with a preliminary analysis covering the shape of our dataset, the attributes data types, and a snippet of our data set. We find that our data had 3513617 rows and 49 columns, 34 of which are categorical variables. Since machine learning algorithms typically do not admit categorical variables, we will have to deal with those during the preprocessing phase.

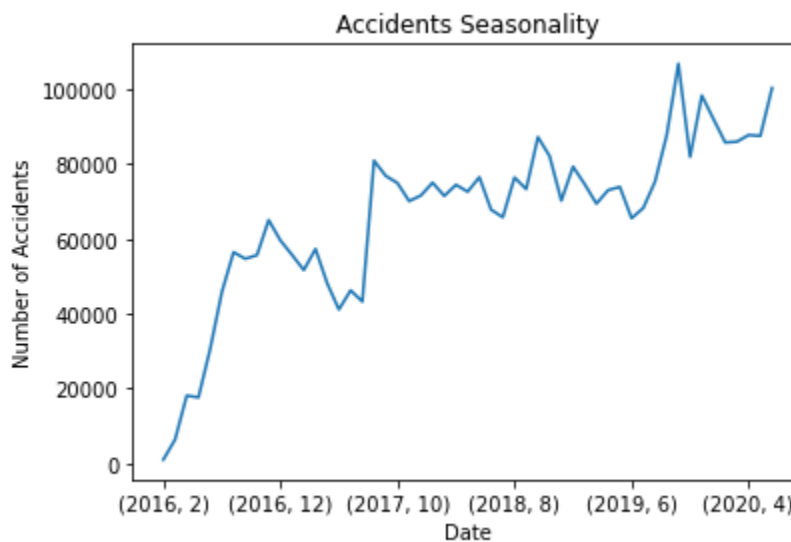
We then begin our exploratory data analysis, starting with the investigation of duplicate rows (we do not find any), and our label variable ‘Severity’ as seen in **Figure 1**.



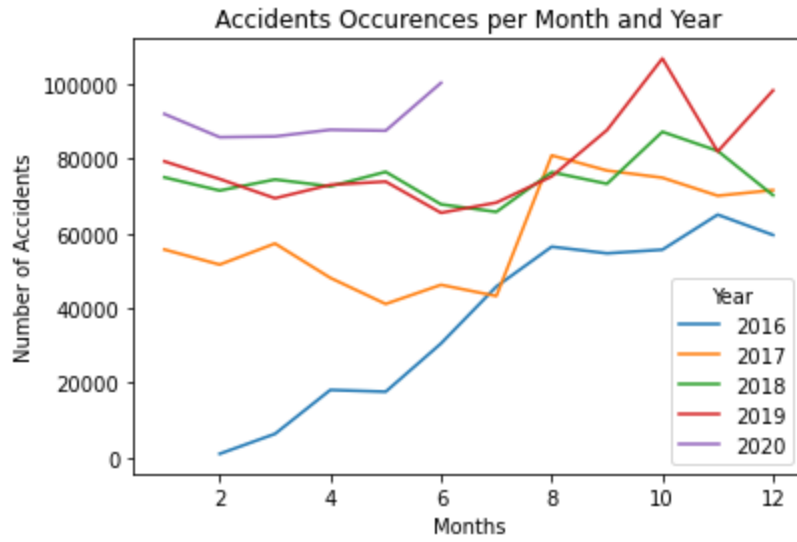
**Figure 1** Target Variable 'Severity'

We notice that our target variable is not balanced such that there are more accidents with severity 2 and 3 as there are with severity 1 and 4. This could induce our machine learning predictive algorithm later on to naively predict all observations severity in the testing set as either 2 and 3, despite the high accuracy on the training set.

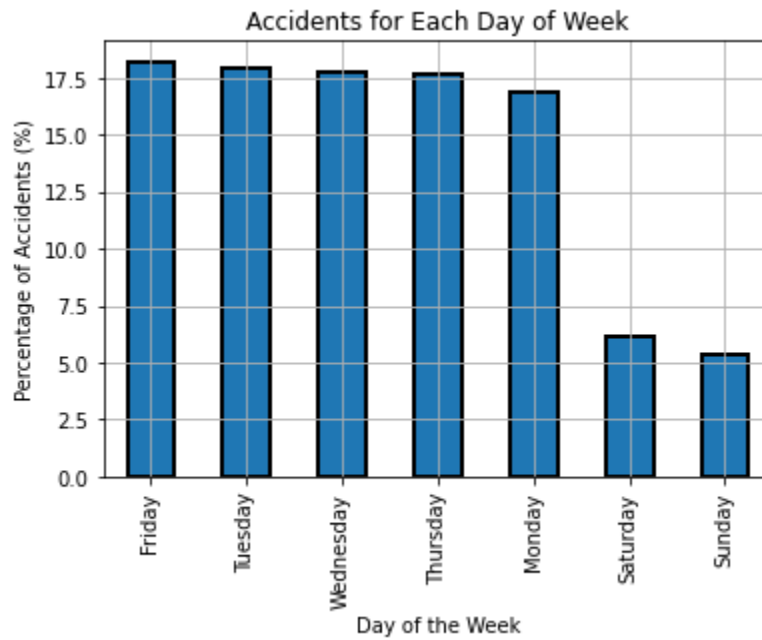
The subsequent analysis of predictors is split into: attributes related to the time of occurrence of accidents, attributes related to location, attributes related to weather, and some miscellaneous investigations of other attributes. We first investigate the seasonality of our data yearly (**Figure 2**), monthly (**Figure 3**), by days of the week (**Figure 4**), and by time of the day (**Figure 5**).



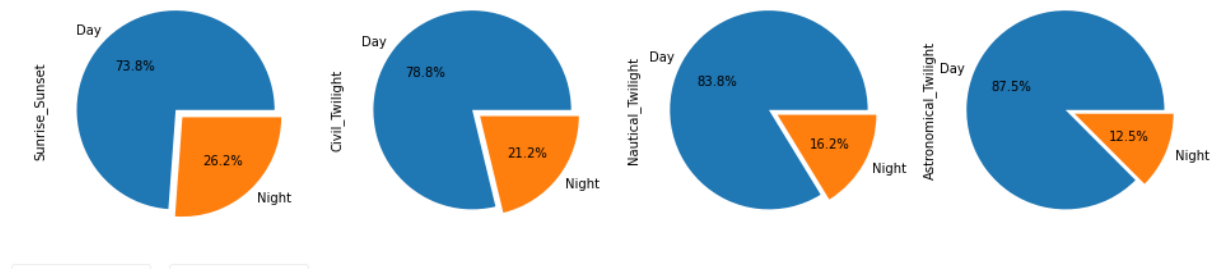
**Figure 2** Yearly Trend of Accidents



**Figure 3** Accidents Seasonality by Year and Month



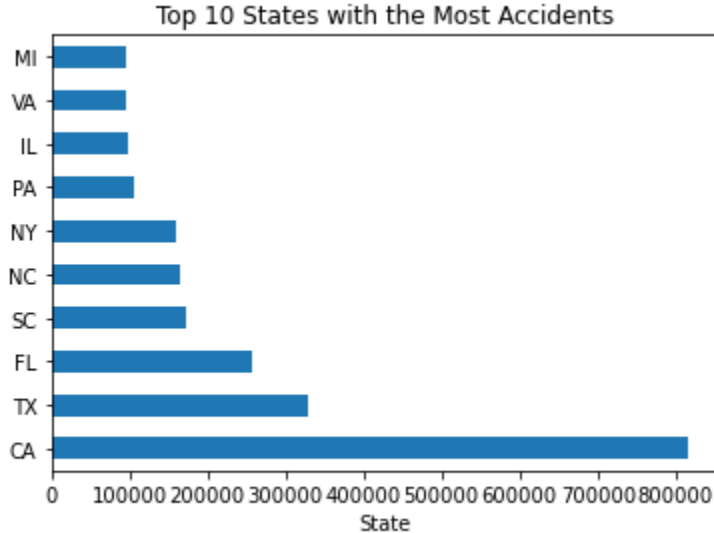
**Figure 4** Accident Seasonality by Day of the Week



**Figure 5** Accidents Occurrence by Timing (Day vs. Night)

From the four figures, we conclude that accidents occurrences in the United States have an upward trend, i.e. they increase from year to year. We also notice that accidents tend to happen more often during the months of September, October, and November. About 90% of accidents are encountered during the weekdays (Monday through Friday) as opposed to 10% on weekends; this may be due to the increased activity and work outings during the weekdays. And finally, accidents tend to occur primarily during the day as seen by the four measures in the last figures.

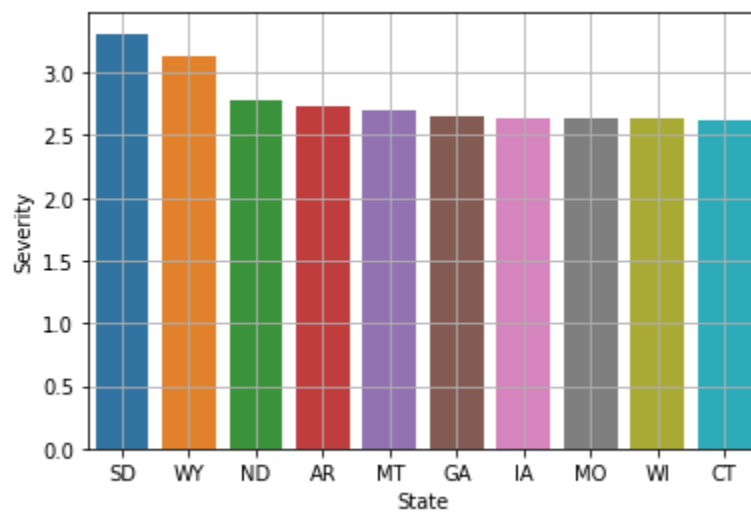
Regarding the location-related attributes, we first investigate accident occurrences by states; we find that the ten states with the most accidents are California (about 23% of the total number of accidents in the data set), followed by Texas, Florida, South Carolina, North Carolina, New York, Pennsylvania, Illinois, Virginia, and Michigan (**Figure 6**).



**Figure 6** The 10 States with the Most Accidents

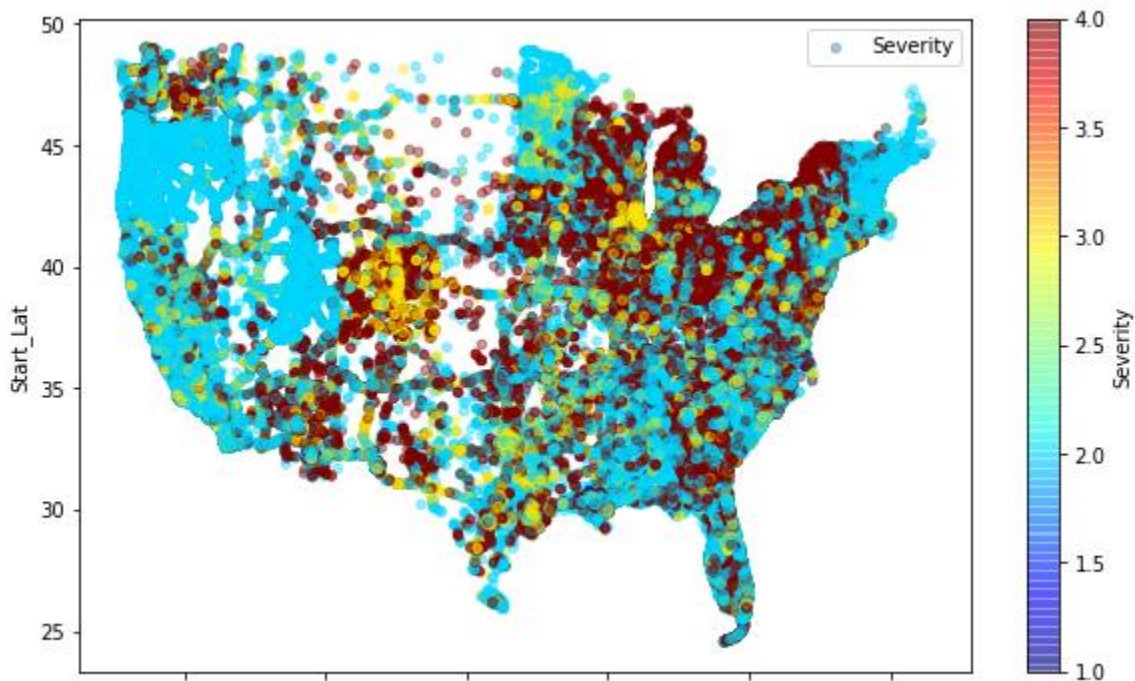
However, the states with the highest average accident severity are South Dakota (average severity higher than 3), Wyoming, North Dakota, Arkansas, Montana, Georgia, Iowa, Missouri,

Wisconsin, and Connecticut (**Figure 7**). That is to say that the states with the most accidents are not necessarily the states with the highest accident severity on average.



**Figure 7** The 10 States with the Highest Accident Severity

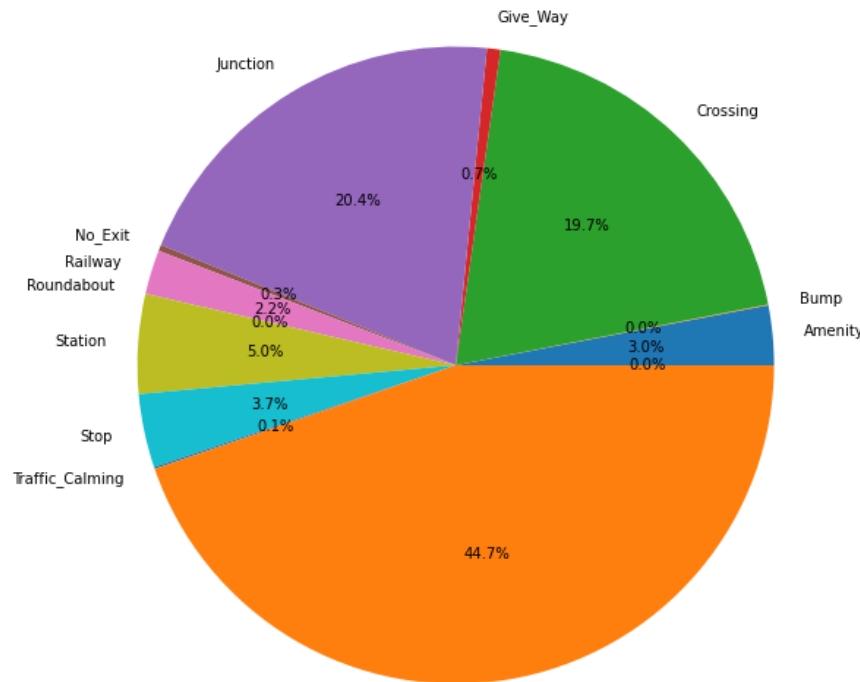
The figure below is a representation of accident severity by geographical coordinates in the United States. The map showcases the results previously obtained, such that for instance South Dakota and Wisconsin are mostly colored with red dots corresponding to high accident severity.



**Figure 8** Accident Severity Across the United States



Finally, in this category of attributes, we look at the impact of traffic objects on accident occurrences. From the pie chart (**Figure 9**), we gather that 44.7% of accidents occur near a traffic signal, 20.4% of accidents occur near junctions, and 19.7% of accidents occur near crossings. That is to say that 84.8% of accidents in our data set occur next to these three traffic objects. Therefore, we keep them as predictors, and we drop the rest of the traffic objects variables.

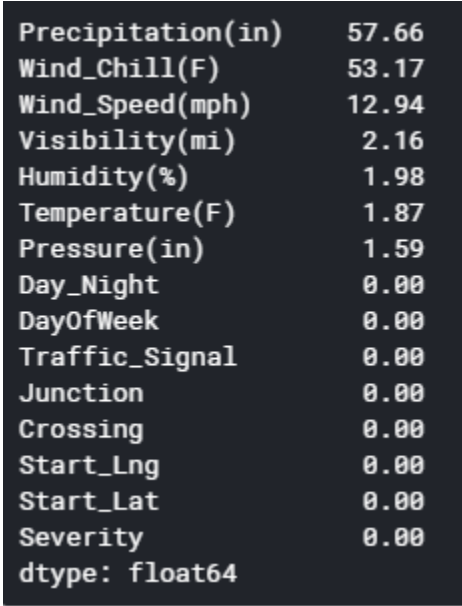


**Figure 9** Accident Occurrences by Traffic Objects

As for the weather conditions during which accidents occur the most often, we find from **Figure 10** that about 39% of accidents occur during clear and fair weather, while only 8% of accidents in our data set occur when it is raining or snowing.



that the accident spans. Then, we investigate the missing values in our data set. As a rule of thumb, we remove the variables with more than 50% of missing data, and for the remaining variables, we replace the missing values with the mean of their respective columns.



Precipitation(in)	57.66
Wind_Chill(F)	53.17
Wind_Speed(mph)	12.94
Visibility(mi)	2.16
Humidity(%)	1.98
Temperature(F)	1.87
Pressure(in)	1.59
Day_Night	0.00
DayOfWeek	0.00
Traffic_Signal	0.00
Junction	0.00
Crossing	0.00
Start_Lng	0.00
Start_Lat	0.00
Severity	0.00
dtype:	float64

**Figure 12** Missing Values

As previously mentioned, since most prediction algorithm cannot work with categorical variables as predictors, we one hot encode our only categorical variable “DayOfWeek” by converting each day of the week into an integer between 0 and 6 inclusive.

Finally, we check correlation between our remaining attributes. As a threshold, any correlation higher than 70% is considered a strong correlation. As seen in **Figure 13**, we do not have any strongly correlated variables. However, we do have a small positive correlation of 45% between Traffic\_Signal and Crossing (as expected, since most crossing are next to a traffic signal), as well as a small negative correlation of 42% between Temperature(F) and Start\_Lat (also expected, since temperature varies by location in the United States).

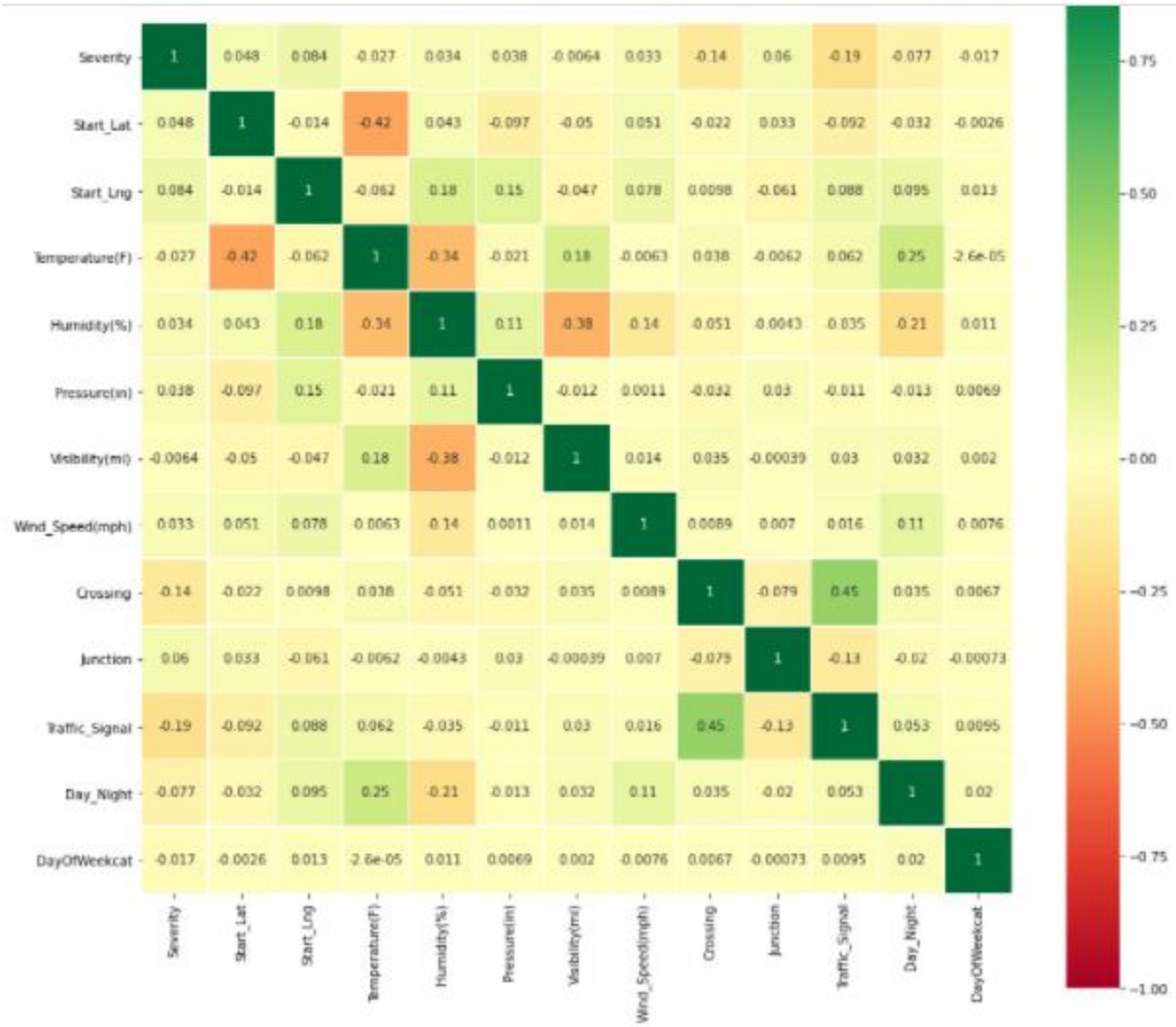


Figure 13 Investigating Correlation

#### 4. Modelling and Predictive Analysis

In this step, we carry out and compare two supervised classification machine learning models, namely: Decision Trees and Logistic Regression. The metrics used for the evaluation are the Jaccard Index, the F1-Score, and the Log Loss.

The Jaccard Index is the ratio of the correctly predicted values to the union of the label test set and prediction set. The F1-Score is a ratio as well involving precision and recall. And the Log Loss is a cost function involving the predicted probability of the output, such that the lower the log loss, the higher the accuracy of the model.

We start our modelling process by splitting our data into a training set (80%) and a testing set (20%). We also normalize our data so that all variables contribute equally to the analysis regardless of their scales.

## Decision Trees Classification

We train a tree of depth 4 with the entropy criterion (the algorithm tries to minimize entropy since it is a measure of randomness and uncertainty). We get a Jaccard Score of 51%, and an F1-Score of 54.5%.

## Logistic Regression

We train a logistic regression model using the solver package *liblinear* which is ideal for large datasets such as ours. We get a Jaccard Score of 51%, an F1-Score of 54.9%, and a Log Loss of 72.9%.

**Table 1** Summary Table of the Evaluation Metrics for all Three Classification Algorithms

Algorithm	Jaccard Score	F1-Score	Log Loss
Decision Trees	51%	54.5%	
Logistic Regression	51%	54.9%	

## **Results & Discussion**

Although the prediction accuracy results are overall relatively low, they are realistic considering that our label target variable is unbalanced.

Further considerations include applying K-Nearest Neighbors and Support Vector Machine algorithms, although they may be slow to run seeing the size of our data set.

## **Conclusion**

All in all, in this paper, we aimed to analyze data pertaining to accident occurrence severity in the United States as a case study. The objective was to predict accident severity to better estimate human injuries and material damages, enable emergency responders to provide tailored and fast care, as well as to avoid traffic congestions. Starting off with our raw data set, we explored and cleaning our data by focusing on removing irrelevant attributes, dealing with missing values, and investigating correlation. Then we carried out on our dataset and compared using different evaluation metrics two classification algorithms, namely: Decision Trees and Logistic Regression. Looking at the accuracy metrics, we concluded that the best classifier is Logistic Regression.

## **References**

[1] OECD. (2017). Road accidents. <https://data.oecd.org/transport/road-accidents.htm>