# Applied Data Science Capstone Project

(Data Science Professional Certificate with IBM)

# Predicting Car Accident Severity

Ibrahim Maassarani

**October 12, 2020**

# Table of Contents

# Table of Figures

## List of Tables

# Chapter 1: Introduction

## 1.1 Background of the Project

Driving cars is a part of our lives in modern societies. It is true that public transportation, as well as alternative transportation methods, are available. However, a large number of people would prefer owning a car since it gives them better control over their time and provides them with much more liberty and autonomy, not to mention that a car is generally faster than a bike for example, especially for longer distances.

Nevertheless, car accidents is a problem that needs to be addressed: it is a major cause of death. For instance, car crashes is a leading cause of death in the United States.

This project aims to design a machine learning based model that is able to predict car accident severity.

## 1.2 Problem Statement

Let us keep in mind that materialistic loss should be avoided as it is important, but it is not the important thing here. This is stated when taking into consideration people's lives that are being lost and the emotional impact of a loved one's death: two losses that are very hard to compensate, if one claims that it is even possible to do so.

Moreover, this is world problem. Therefore, it is necessary to narrow down the scope of the project so it becomes feasible to deliver required results in time.

In this project, we aim to leverage the power of Machine Learning, more specifically classification algorithms (KNN, SVM, Random Forest…) since we are facing a classification problem, to predict cars accident severity, based on the data in hand.

We will be focusing on the Emerald City, i.e., Seattle, United States. Hence, our project is mainly aimed at predicting the severity of a car accident in Seattle City.

## 1.3 Significance of the Project

Obviously, the problem we are facing is a serious one. We believe that our work will offer insights on how to drive safely. Also, it will help responsible parties in taking necessary actions: take more precautions in certain locations where collisions might be more frequent for example, or reduce the speed limit in certain areas, or even handle a current collision, etc…

This will ensure that the severity of crashes, as well as their number, are reduced, all of which means that there will be less property and other materialistic damage, and what is much more important, less serious injuries and human casualties.

# Chapter 2: Data

## 2.1 Data Sources

You can find the dataset mentioned in the Capstone guidelines here. We first used it, but then we found a newer version on Kaggle that was more organized and had more observations. Therefore, we used the dataset found on Kaggle.

You can click here for a map showing the distribution of the crashes, an overview of the data, as well as table representation of the data itself.

Moreover, you can also click here for a summary of the dataset, available in a PDF file. This summary is very important as it helps us to better understand our data, i.e., the available fields and what they represent, as well as their possible values and a description of each value.

Also, the geopy python package was used to get Seattle’s coordinates (longitude and latitude).

## 2.2 Data Description

The downloaded data, a CSV file, was transformed into one Pandas dataframe. It initially consisted of a total of 40 columns (fields) and 221738 entries (observations).

Based on the problem description, some factors will influence our results, and others will not:

- The location will most probably affect an accident's severity: it is intuitive that the number of crashes, as well as their severity, in crowded areas will be higher than in the uncrowded areas.
- The junction type (intersection, driveway, mid-block...) probably has an effect.
- There are more than one identifier field that we tend to believe to be irrelevant to our work, and we shall investigate our hypothesis while exploring the data.
- We shall see if a certain month or day of week witnesses more accidents than others. If such correlation fail to exist, it will most probably have no effect on a crash's severity. This can be rechecked after deploying our model and testing its performance: we could reprocess the data, make changes in our feature selection, etc...

We could take advantage of many available attributes: PERSONCOUNT to see if the number of persons in an accident affects the accident's severity, as well as other attributes that we believe it will affect an accident's severity such as WEATHER, ROADCOND, LIGHTCOND, etc...

Detailed Description is available in the link given in the summary mentioned in (2.1).

## 2.3 Data Cleaning

Thankfully, there was no duplicated data.

However, there are many problems with our dataset. Many fields hold lots of missing values (NaN, Other, sometimes 0) as shown in the notebook after taking a look at the dataframe's info. Hence, thorough investigation was needed. Many fields, as well as certain rows, were dropped as explained in (2.4).

SEVERITYCODE denotes the severity of an accident, i.e., what we wish to predict. It contained one null value, so this null value was dropped. Also, when SEVERITYCODE is 0, it means that the accident is of 'Unknown Severity'. Therefore, this is considered as missing data for the class we are trying to predict. Hence, all entries with SEVERITYCODE 0 were dropped. What we did also solved the problem where 21654 values out of 21656 values in the WEATHER field were missing. These values corresponded to SEVERITYCODE 0: they were dropped when all rows with SEVERITYCODE 0 were dropped.

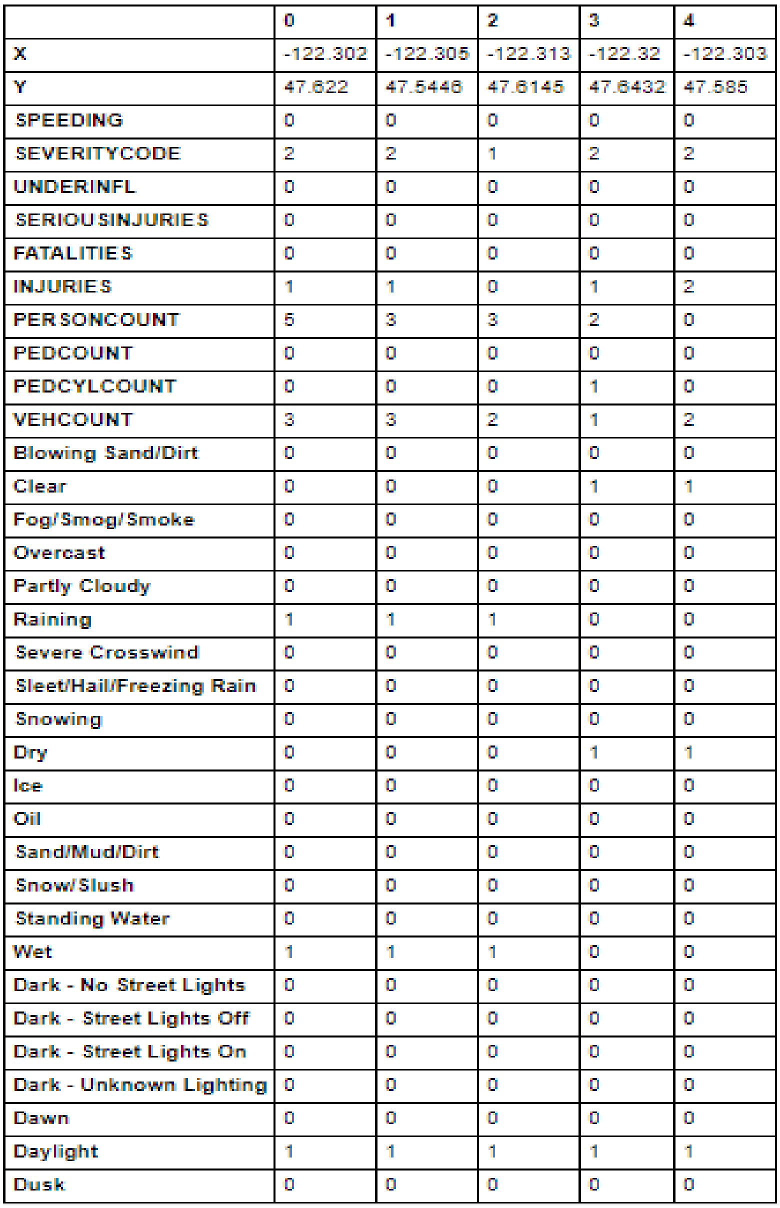
SPEEDING indicates whether speed played a role in the accident or not. If yes than it is a 'Y', and 'NaN' if otherwise (no). This explains why there was so much missing values in SPEEDING. This was handled by mapping SPEEDING and changing the 'Y' to '1' and 'NaN' to '0': 1 and 0 were chosen to be integers to represent a Boolean category.

UNDERINFL indicates whether the driver was under the influence of alcohol or drugs: '0' and 'N' if no, and '1' and 'Y' if yes.  This is not as in SPEEDING. We still need to change the 1s and 0s because they are strings, and we need them to be integers for the same reason as in SPEEDING. Hence, 'Y' and '1' (str type object) were changed to '1' (int type object), and 'N' and '0' (str type object) were changed to '0' (int type object).

After drawing a heat map of missing records, and calculating the number of rows with missing values w.r.t. the number of total rows, it was found to be approximately 19.23%. At first sight, it seems that it will have a great influence on our data: we are considering to drop one fifth of our data. However, we have 221738 records, i.e., we have sufficient data for our machine learning algorithms to accurately predict the class of each observation. This claim was validated after evaluating our models in (4.2) where they all had excellent performance.

Not to forget that we encoded (one-hot encoding) all of the categorical data, i.e., WEATHER, ROADCOND, and LIGHTCOND.

We are now dealing with a dataset of 15 columns (features) and 171504 records of consistent data.



| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| X | -122.302 | -122.305 | -122.313 | -122.32 | -122.303 |
| Y | 47.622 | 47.5446 | 47.6145 | 47.6432 | 47.585 |
| SPEEDING | 0 | 0 | 0 | 0 | 0 |
| SEVERITYCODE | 2 | 2 | 1 | 2 | 2 |
| UNDERINFL | 0 | 0 | 0 | 0 | 0 |
| SERIOUSINJURIES | 0 | 0 | 0 | 0 | 0 |
| FATALITIES | 0 | 0 | 0 | 0 | 0 |
| INJURIES | 1 | 1 | 0 | 1 | 2 |
| PERSONCOUNT | 5 | 3 | 3 | 2 | 0 |
| PEDCOUNT | 0 | 0 | 0 | 0 | 0 |
| PEDCYLCOUNT | 0 | 0 | 0 | 1 | 0 |
| VEHCOUNT | 3 | 3 | 2 | 1 | 2 |
| Blowing Sand/Dirt | 0 | 0 | 0 | 0 | 0 |
| Clear | 0 | 0 | 0 | 1 | 1 |
| Fog/Smog/Smoke | 0 | 0 | 0 | 0 | 0 |
| Overcast | 0 | 0 | 0 | 0 | 0 |
| Partly Cloudy | 0 | 0 | 0 | 0 | 0 |
| Raining | 1 | 1 | 1 | 0 | 0 |
| Severe Crosswind | 0 | 0 | 0 | 0 | 0 |
| Sleet/Hail/Freezing Rain | 0 | 0 | 0 | 0 | 0 |
| Snowing | 0 | 0 | 0 | 0 | 0 |
| Dry | 0 | 0 | 0 | 1 | 1 |
| Ice | 0 | 0 | 0 | 0 | 0 |
| Oil | 0 | 0 | 0 | 0 | 0 |
| Sand/Mud/Dirt | 0 | 0 | 0 | 0 | 0 |
| Snow/Slush | 0 | 0 | 0 | 0 | 0 |
| Standing Water | 0 | 0 | 0 | 0 | 0 |
| Wet | 1 | 1 | 1 | 0 | 0 |
| Dark - No Street Lights | 0 | 0 | 0 | 0 | 0 |
| Dark - Street Lights Off | 0 | 0 | 0 | 0 | 0 |
| Dark - Street Lights On | 0 | 0 | 0 | 0 | 0 |
| Dark - Unknown Lighting | 0 | 0 | 0 | 0 | 0 |
| Dawn | 0 | 0 | 0 | 0 | 0 |
| Daylight | 1 | 1 | 1 | 1 | 1 |
| Dusk | 0 | 0 | 0 | 0 | 0 |

*Table 1: 5 Samples of the Cleaned Dataset (prior to scaling)*

Last but not least, we shuffled, standardized our data, from which we constructed the features dataset and the class dataset. The features and class datasets were then split into a training dataset and a testing dataset on which each model will be trained and tested, respectively.

Finally, no more work was needed after training and testing our models, for which the reason will be shown in (4.2).

## 2.4 Feature Selection

Thanks to the map that was drawn in our notebook, it was obvious that the location had a strong impact on the number of accidents, and most probably on an accident severity. The location was represented in X and Y fields (longitude and latitude of the crash).

JUNCTIONYPE does have a location. Hence, we tend to believe that the impact of a junction will be embedded in the location, i.e., the coordinates. Therefore, there is no need to take JUNCTIONTYPE into consideration (the whole column was dropped).

SEVERITYCODE denotes a collision's severity. It is the class that we aim to predict. It is very clear that we need this field.

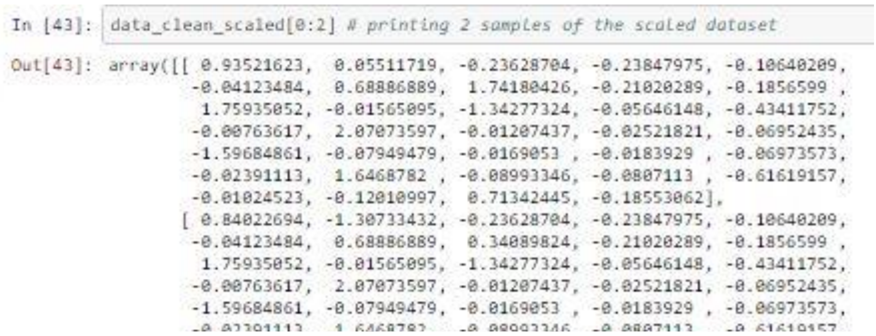
Certain fields obviously causes car accidents and affect their severity such as ROADCOND (road condition), LIGHTCOND (light condition), WEATHER (weather condition), and others.

Further investigation showed that the date and time of an accident had weak or insignificant correlation with crashes severity, so INCDATE and INCDTTM that denoted the date and time of a collision were both dropped.

Also, the dataset holds many attributes that are irrelevant in our analysis. For instance, OBJECTID, INCKEY and REPORTNO, as well as other fields, are clearly identifiers, which indicates that the csv file was imported from a database table. Other fields, such as LOCATION and SDOTLCODE, as well as others, and their descriptions, if any, holds categorical data with many different categories. Hence, these fields will disrupt our analysis. Therefore, all such fields were dropped.

At the end, 15 features were selected: X, Y, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, SEVERITYCODE, UNDERINFL, SERIOUSINJURIES, FATALATIES, INJURIES, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, and VEHCOUNT.

As our designed models performed excellently, we could strongly claim that our features were chosen effectively.



```
In [43]: data_clean_scaled[0:2] # printing 2 samples of the scaled dataset

Out[43]: array([[ 0.93521623,  0.05511719, -0.23628704, -0.23847975, -0.10640209,
                 -0.04123484,  0.68886889,  1.74180426, -0.21020289, -0.1856599 ,
                  1.75935052, -0.01565095, -1.34277324, -0.05646148, -0.43411752,
                 -0.00763617,  2.07073597, -0.01207437, -0.02521821, -0.06952435,
                 -1.59684861, -0.07949479, -0.0169053 , -0.0183929 , -0.06973573,
                 -0.02391113,  1.6468782 , -0.08993346, -0.0807113 , -0.61619157,
                 -0.01024523, -0.12010997,  0.71342445, -0.18553062],
                [ 0.84022694, -1.30733432, -0.23628704, -0.23847975, -0.10640209,
                 -0.04123484,  0.68886889,  0.34089824, -0.21020289, -0.1856599 ,
                  1.75935052, -0.01565095, -1.34277324, -0.05646148, -0.43411752,
                 -0.00763617,  2.07073597, -0.01207437, -0.02521821, -0.06952435,
                 -1.59684861, -0.07949479, -0.0169053 , -0.0183929 , -0.06973573,
                 -0.02391113,  1.6468782 , -0.08993346, -0.0807113 , -0.61619157
```

*Figure 1: 2 Samples of the Scaled Dataset*

# Chapter 3: Methodology

## 3.1 Exploratory Data Analysis

### 3.1.1 Car Crashes Number vs. Location

As stated earlier, location is describe by both X and Y fields, i.e., longitude and latitude. In the map shown in Figure 2, we are visualizing the first 1000 collisions of our dataset that do not contain null values. It is clear that the number of accidents differ from one region to another. Hence, the location of an accident is related to our project's aim. It is purely intuitive that most probably an accident's severity on a freeway is much higher than on a cobblestone street.

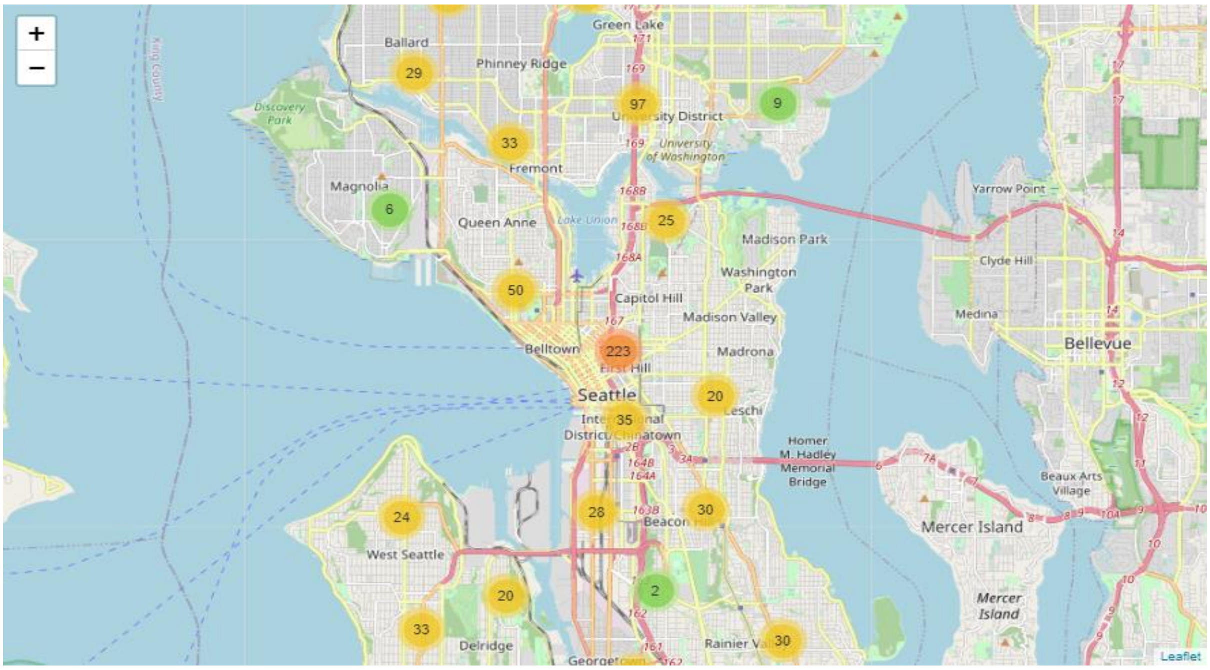


*Figure 2: Seattle Map showing Accidents Distribution by Location*

### 3.1.2 Car Crashes Number vs. Month / Day of Week by Severity

We were trying to see if there was a correlation between time, specifically a certain month or day of week, and the number of crashes for each severity.

From Figure 3, we can notice the following:

- Severity 1: the accidents count is almost equally distributed throughout the year
- Severity 2: accidents are more frequent during summer and fall
- Severity 2b: accidents are the highest during the summer
- Severity 0: Accidents are a little more frequent as the end of the year approaches
- Severity 3: Accidents are less frequent during winter and early summer

The distribution for the accidents number of all severities almost look the same, except for where Severity is 3. It is a good feature to weaken the probability that the accident has a Severity 3 during winter or early summer. However, let us not forget that in Severity 3, we have casualties, which all alone is an enough feature (no casualties => not Severity 3). Hence, there is no need to add a MONTH field.

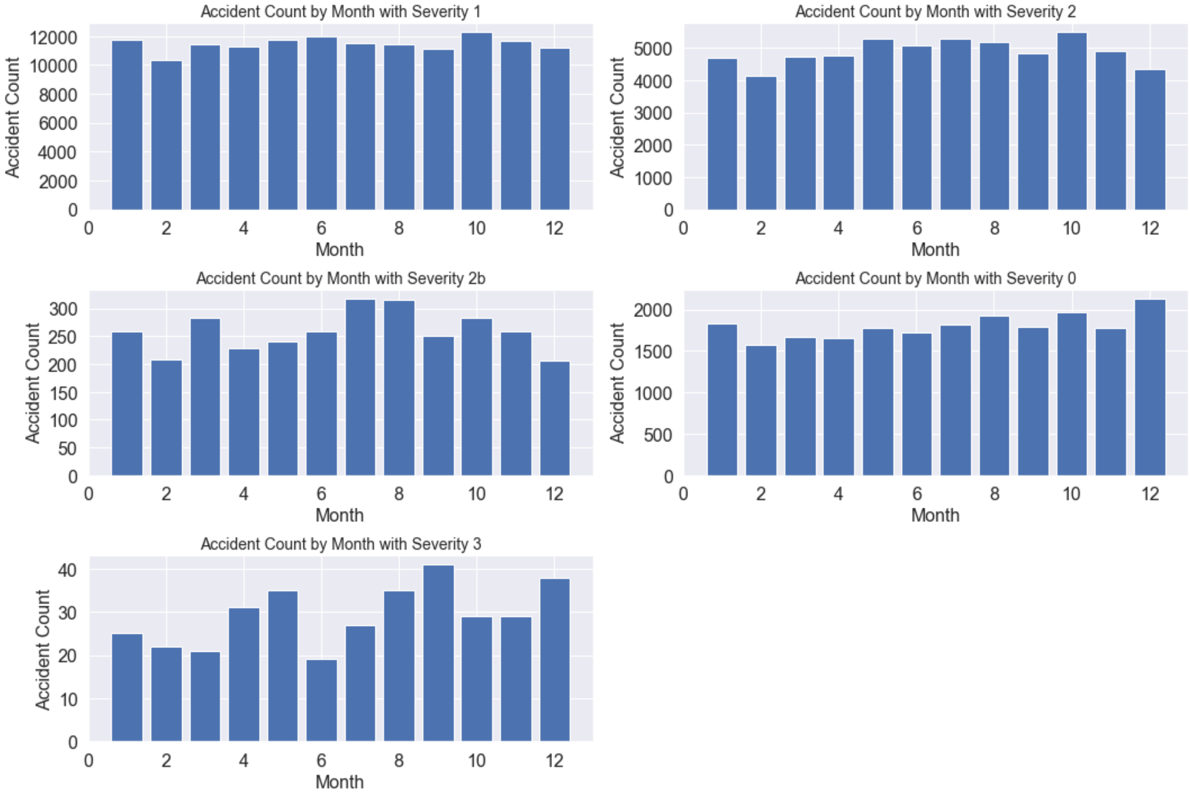


*Figure 3: Total Monthly Accidents by Severity*

In Figure 4, the first day of the week (0) is Monday. We can notice that there is an important drop in the count during weekends (5 and 6, i.e., Saturday and Sunday) across all severities. Moreover, we can see that it is not the case in Severity 3.

Nevertheless, we could say that the attribute we have created, DAYOFWEEK is not of great importance to predict an accident's severity since 4 out of 5 severities have very similar distributions with respect to the day of the weeks (same matter explained earlier with Month regarding Severity 3).

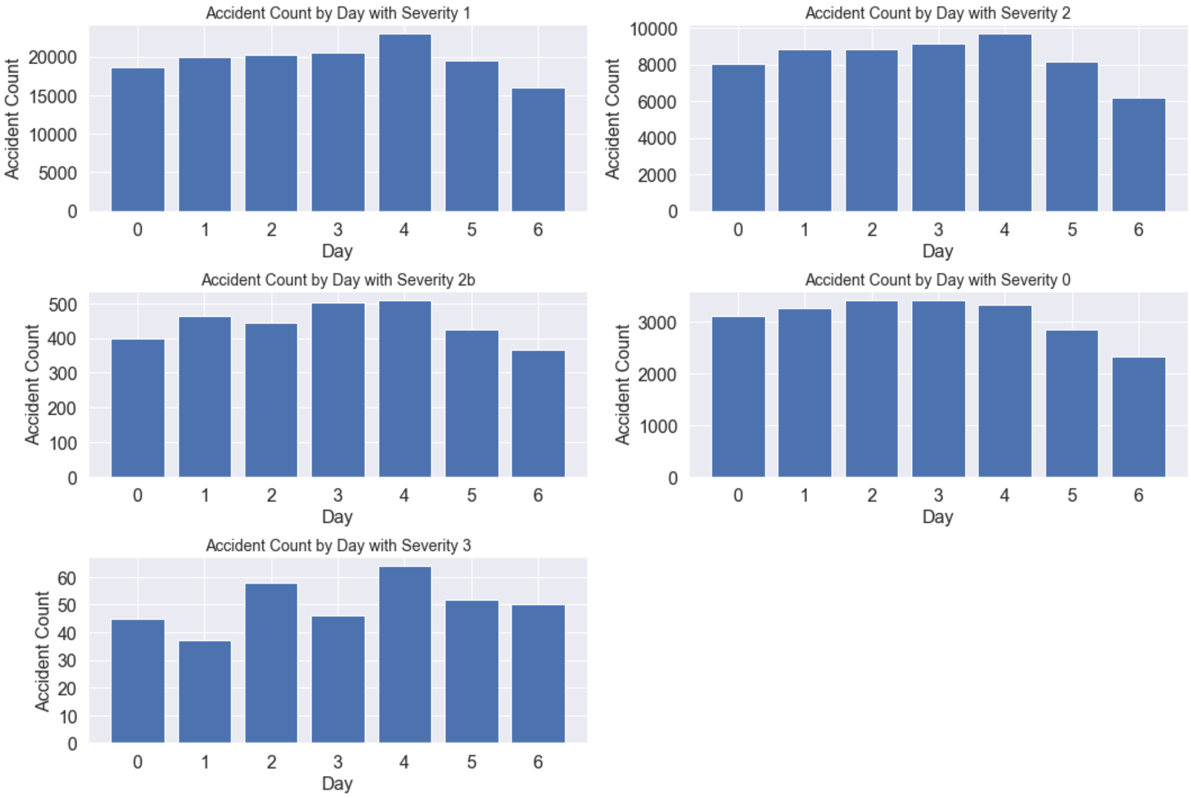
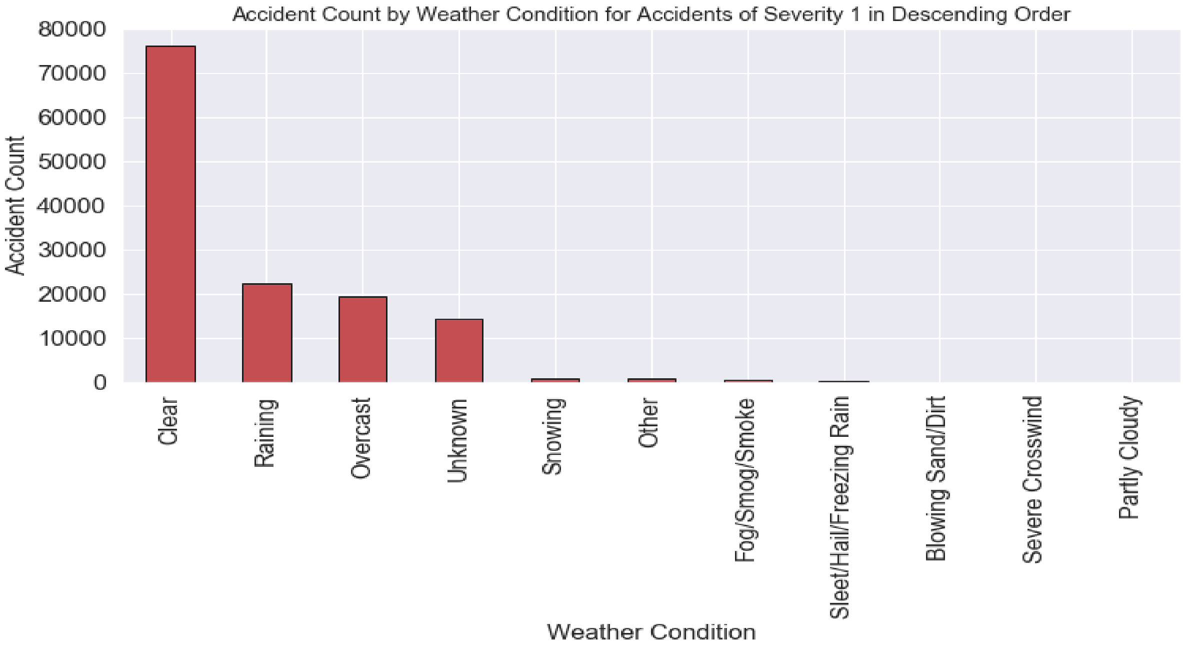
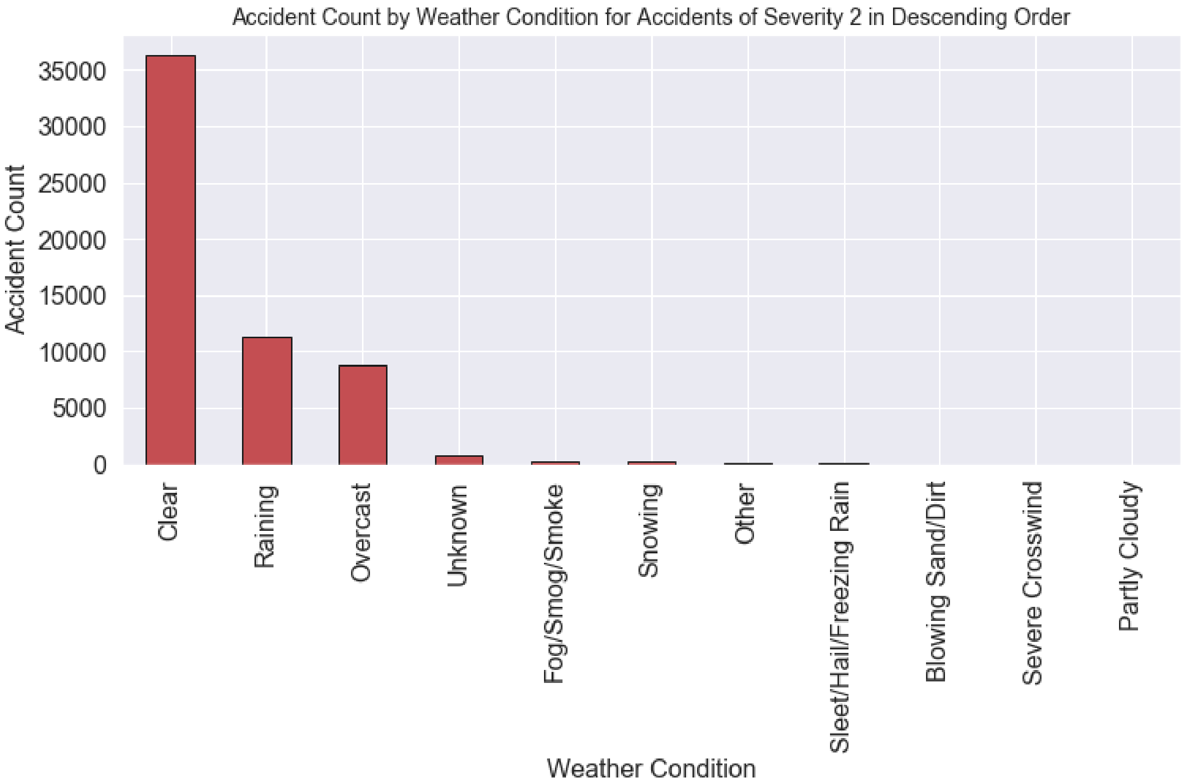


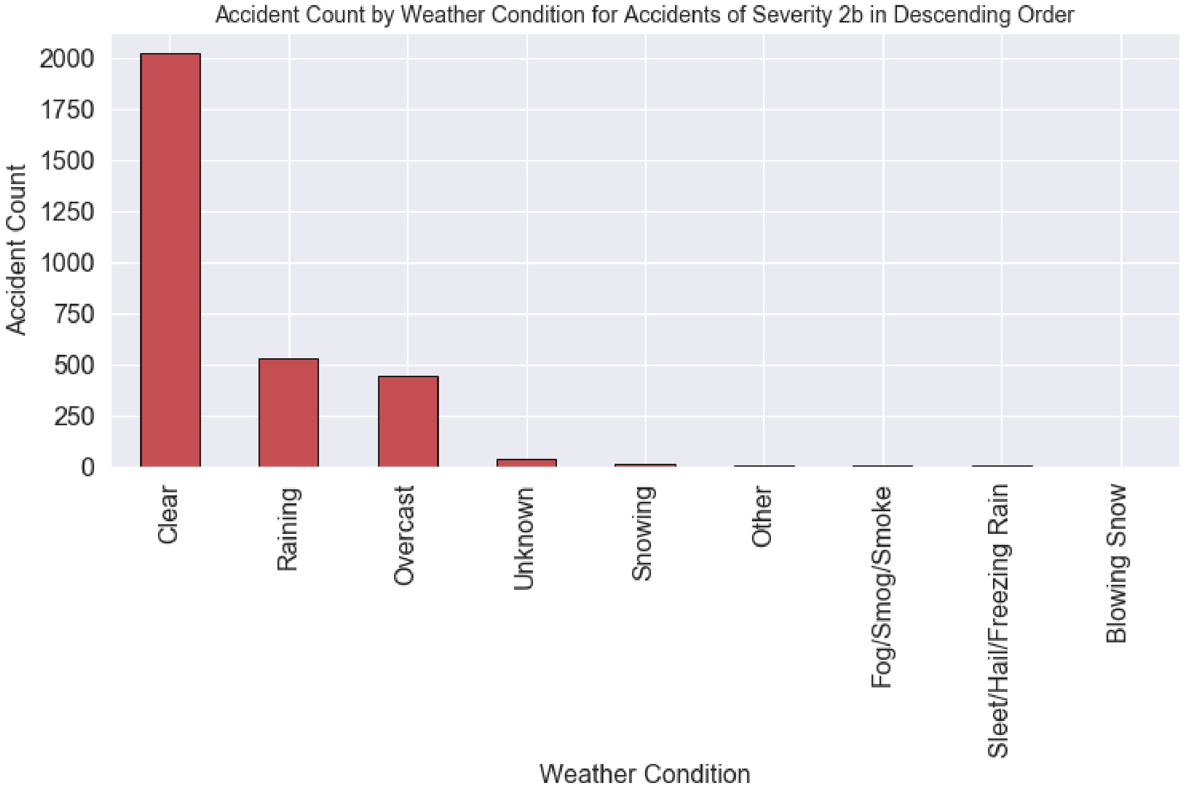
*Figure 4: Total Daily Accidents by Severity*

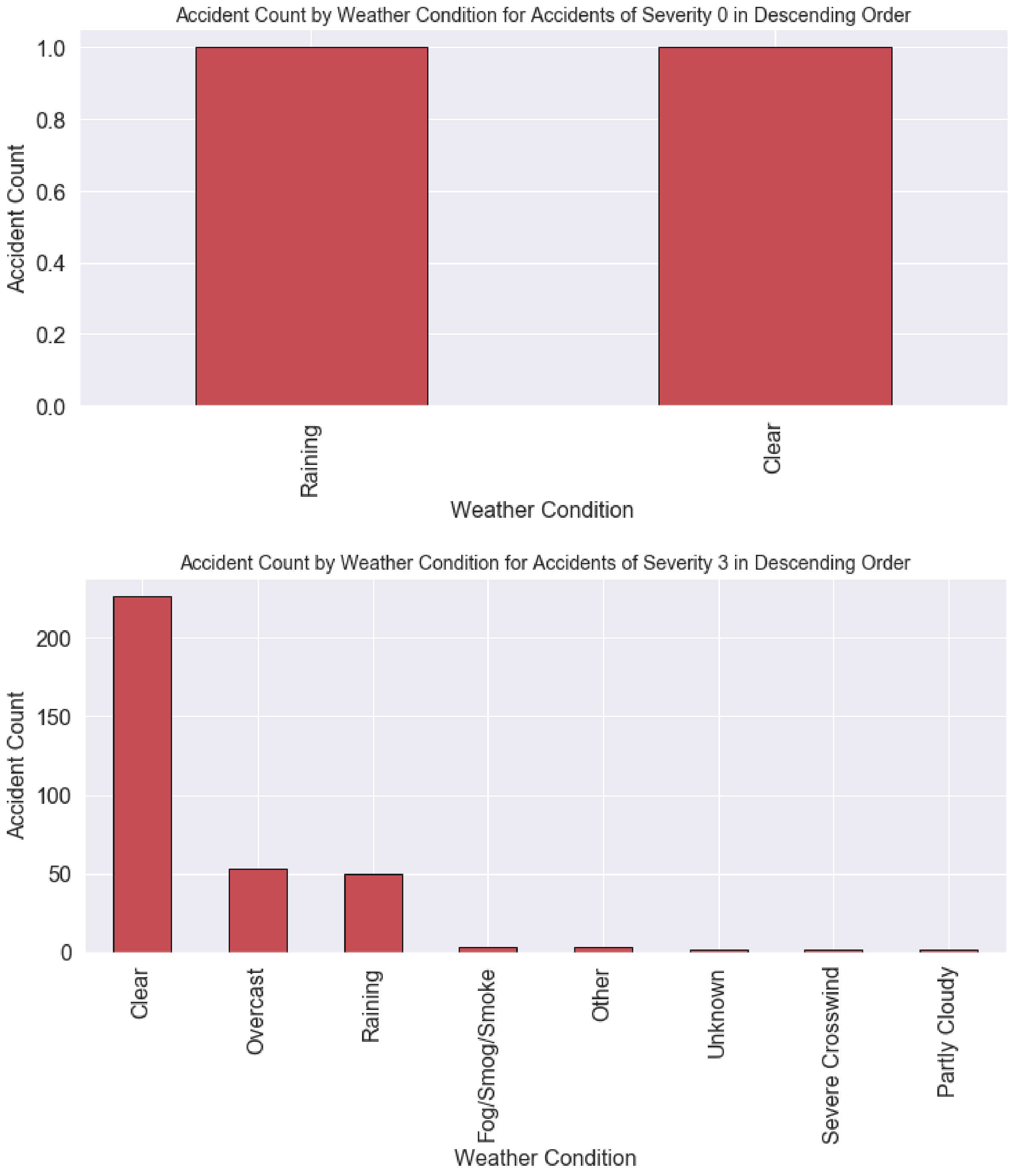
### 3.1.3 Car Crashes Number vs. Weather by Severity

As for the weather conditions, we can see from Figure 5 that for all severities, the majority of accidents occurs under Clear, Raining and Overcast Weather. It is true that there as an important number of accidents where Severity is 1 and the weather condition is unknown, but it does not break our general conclusion.




Accident Count by Weather Condition for Accidents of Severity 2 in Descending Order
Accident Count
35000
30000
25000
20000
15000
10000
5000
0
Clear
Raining
Overcast
Unknown
Fog/Smog/Smoke
Snowing
Other
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
Weather Condition


Accident Count by Weather Condition for Accidents of Severity 2b in Descending Order
Accident Count
2000
1750
1500
1250
1000
750
500
250
0
Clear
Raining
Overcast
Unknown
Snowing
Other
Fog/Smog/Smoke
Sleet/Hail/Freezing Rain
Blowing Snow
Weather Condition



Accident Count by Weather Condition for Accidents of Severity 0 in Descending Order

Accident Count by Weather Condition for Accidents of Severity 3 in Descending Order

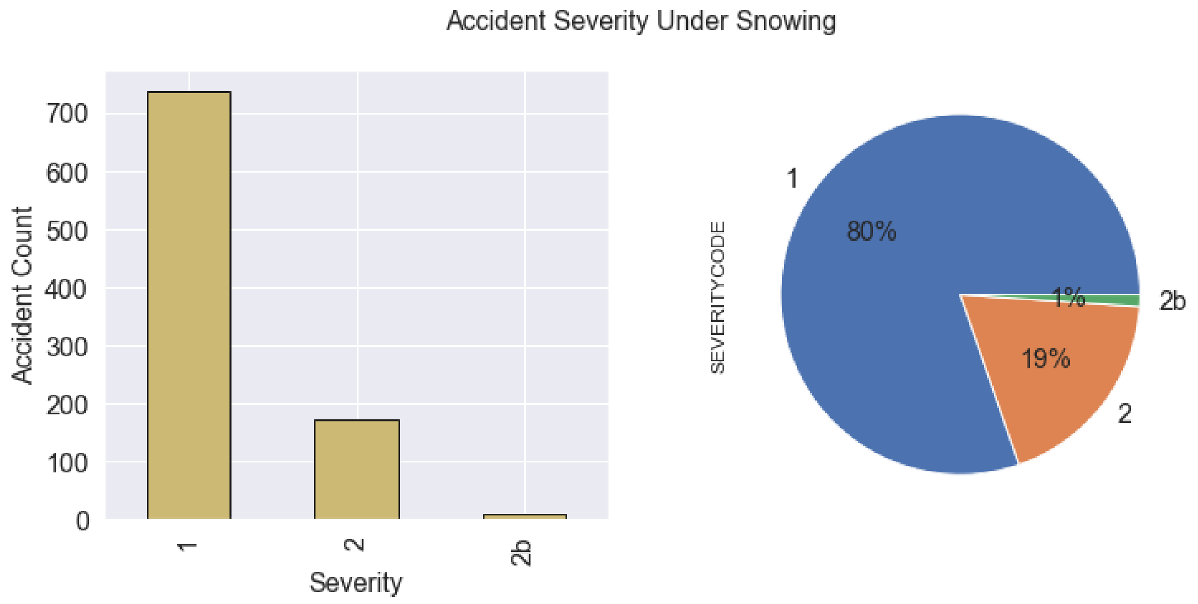
*Figure 5: Accidents Number by Severity, by Weather*

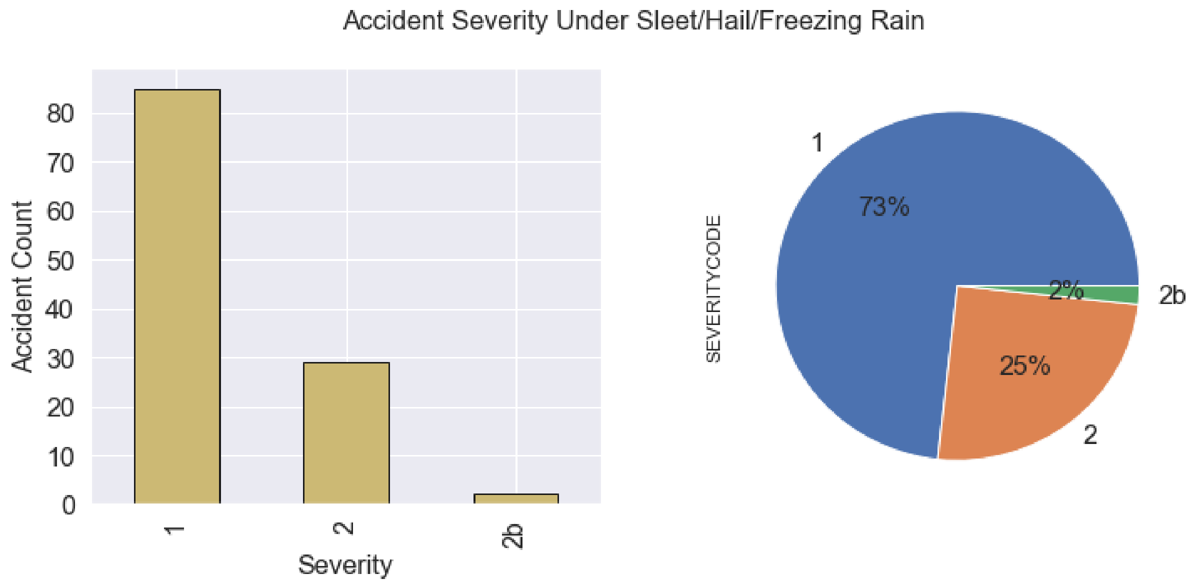
We can also conclude that the weather condition is generally irrelevant to the accident severity. However, there are special weather conditions that are worth investigating such as Snowing and Fog/Smog/Smoke, as well as similar conditions, that most probably affect an accident's severity (slippery road, unclear vision...). One point worth mentioning is that the weather conditions are not identically ordered in all of the graphs: this could be of some significance. Still, further investigation is needed to be done to determine whether WEATHER is going to be selected or not.

Nevertheless, one might find an inconsistency in the graphs. In Severity 0, there are only 2 accidents. This issue was explained in (2.3) when we talked about Severity 0: for all accidents of Severity 0, except for 2 accidents, weather conditions were unknown. But all of the corresponding values (available and missing) of Severity 0 were dropped since every record with Severity 0 was dropped.

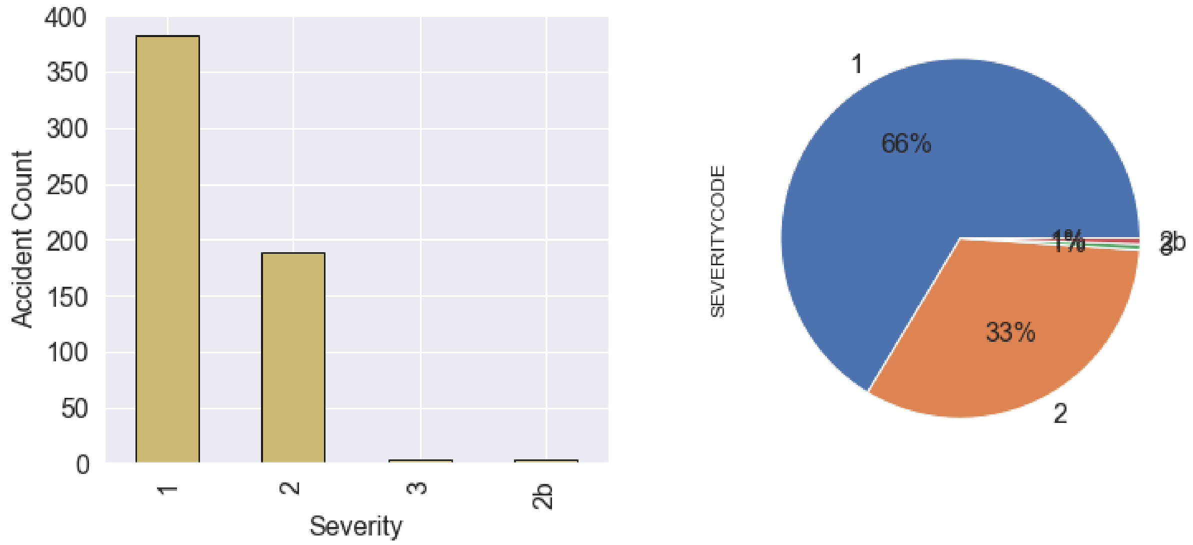
### 3.1.4 Car Crashes Number vs. Special Weather Conditions by Severity

As for the weather conditions, we can see from Figure 6 that for all severities, the majority of accidents occurs under Clear, Raining and Overcast Weather. It is true that there as an important number of accidents where Severity is 1 and the weather condition is unknown, but it does not break our general conclusion.

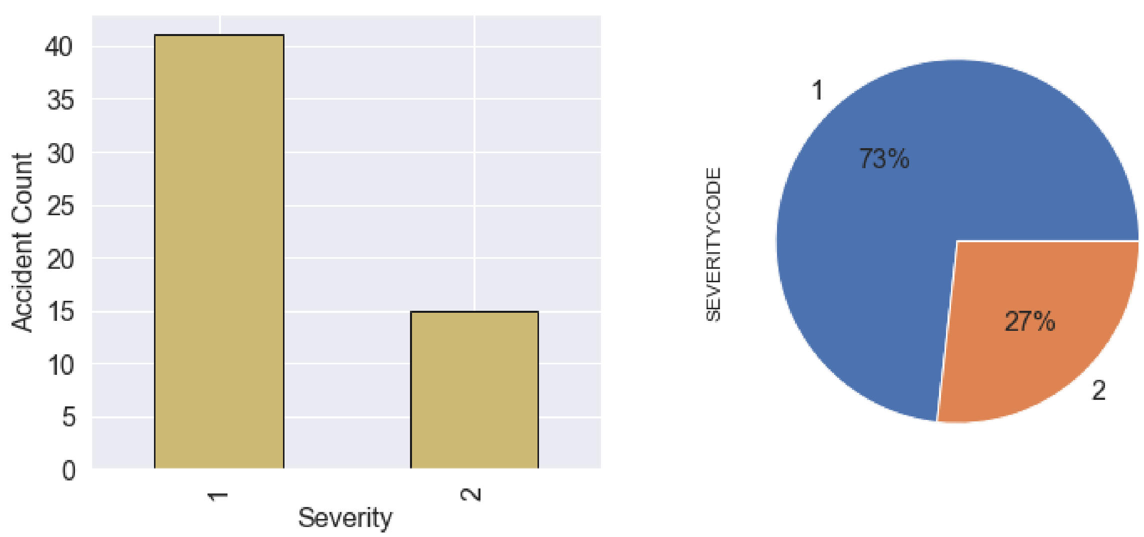




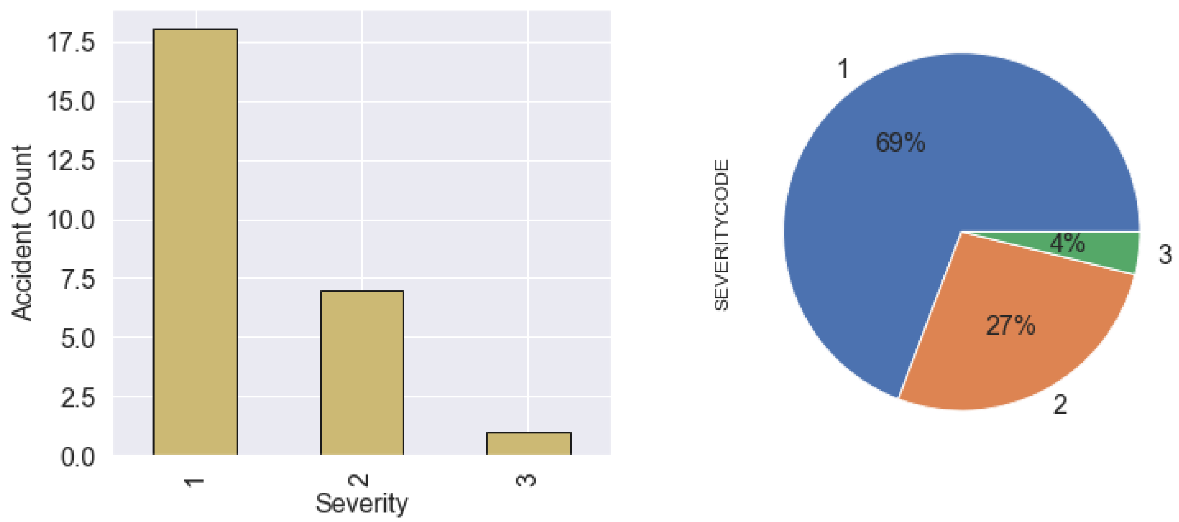
Accident Severity Under Fog/Smog/Smoke

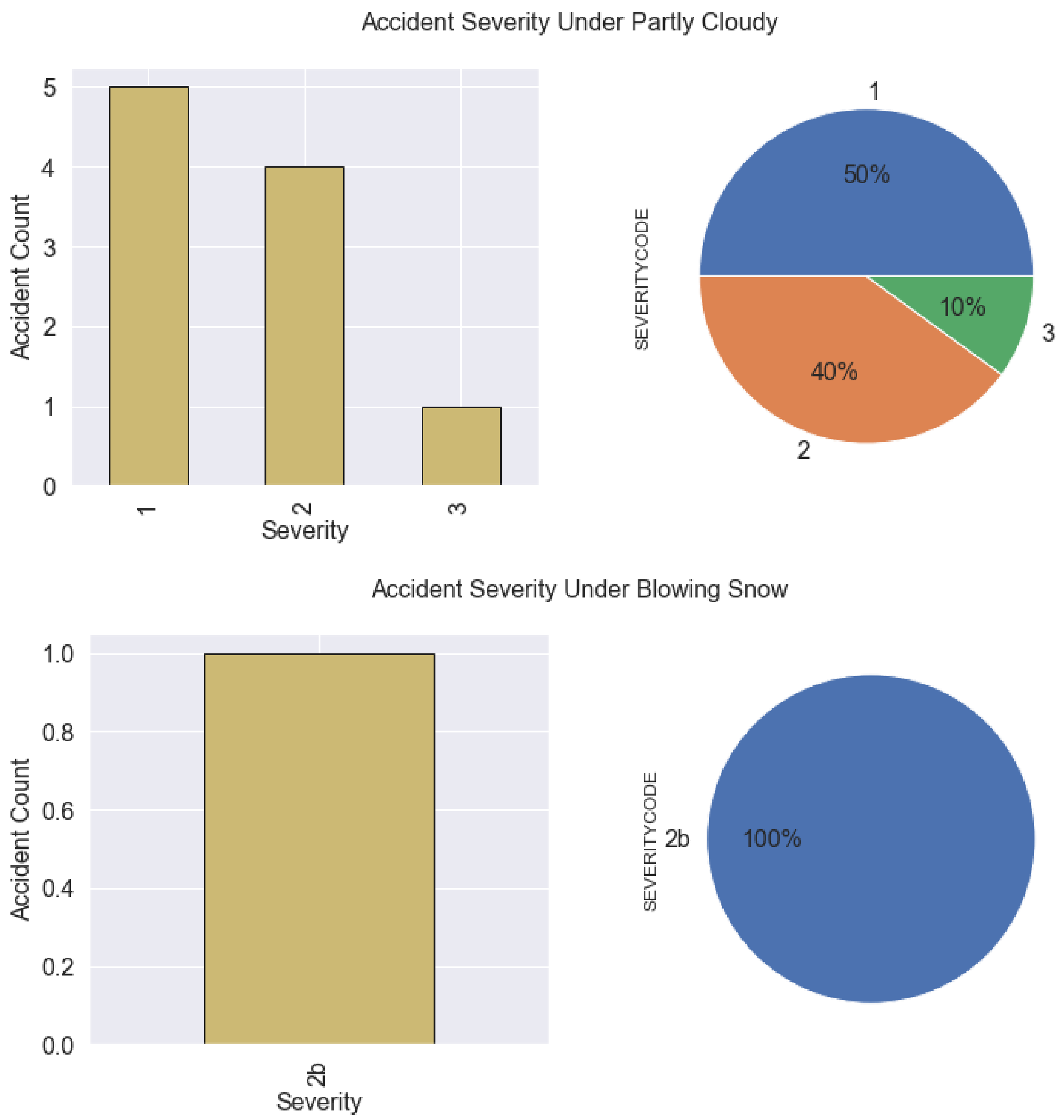


Accident Severity Under Blowing Sand/Dirt



Accident Severity Under Severe Crosswind




Accident Severity Under Partly Cloudy

Accident Severity Under Blowing Snow

*Figure 6: Accidents Number and Partition by Severity, by Special Weather Condition*

It seems that there is a correlation between these conditions and the accident severity. For example, Severity 2 increased from 19% (under Snowing) to 40% (under Partly Cloudy), Severity 1 decreased from 80% to 50% and Severity 2b increased from 1% to 10% (under the same conditions, respectively). Another example is that under Blowing Snow, Severity 2b is 100%: this is an important mark for this Severity Code.

### 3.1.5 Car Crashes Number vs. Speeding by Severity

Speeding indicates whether speed played a role in the accident or not. If there was speeding, it is a 'Y' (yes), and 'NaN' if otherwise (no).

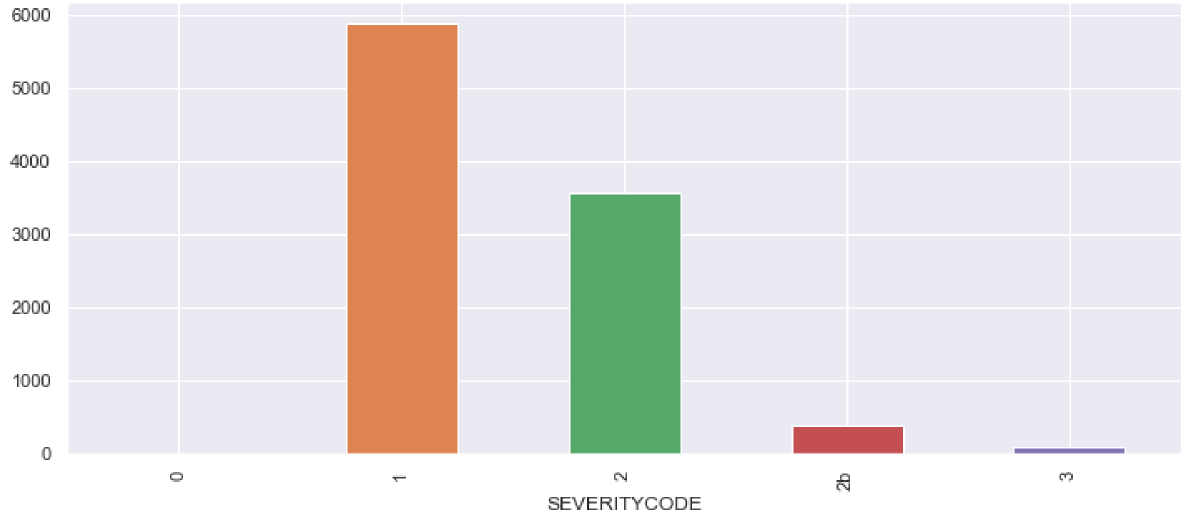


*Figure 7: Total Cases where Speeding Played a Role, by Severity*

### 3.1.5 Correlation between Selected Features in the Clean Dataset

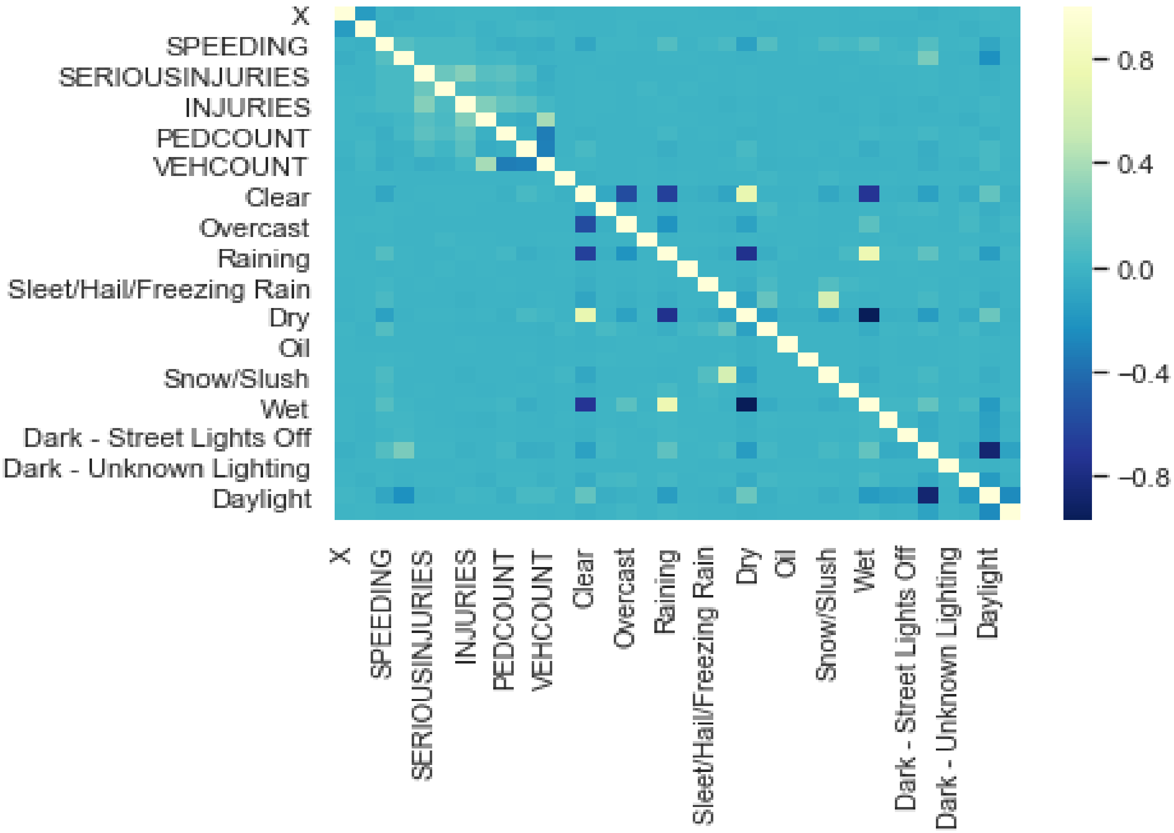


*Figure 8: Heat Map Showing Correlation between Different Features*

It is obvious from the heat map in Figure 8 that certain features do have strong positive or negative correlation, while the majority has a weak or no correlation at all.

## 3.2 Machine Learning Model Selection

The output variable is a label / category / class that we wish to predict based on observed values. Our dataset already contains the classes for each observation, i.e., our data is pre-labeled. Hence, this is a traditional classification problem. This is why classical classification algorithms were used: Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest.

There are more algorithms out there, but testing and evaluating 4 algorithms shall suffice for the scope of our work.

# Chapter 4: Results

Results will be discussed while displaying the classification report of each model, as well as the confusion matrix.

## 4.1 Logistic Regression

### 4.1.1 Confusion Matrix

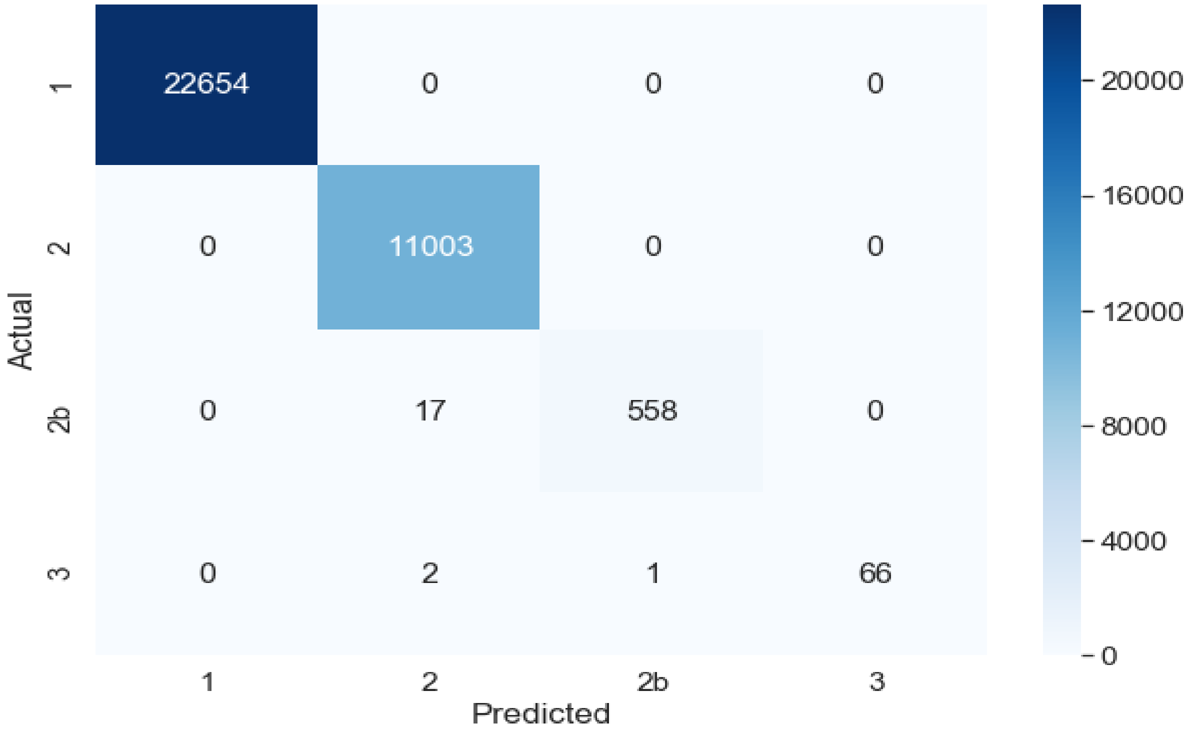


*Figure 9: Logistic Regression Confusion Matrix*

The way to look at is the following:

1- For the actual values, we calculate the sum of every row:

- Severity 1: 22654 + 0 + 0 + 0 = 22654 => 22654 actual crashes of Severity 1
- Severity 2: 11003 actual crashes of Severity 2
- Severity 2b: 0 + 17 + 558 + 0 = 575 actual crashes
- Severity 3: 0 + 2 + 1 + 66 = 69 actual crashes

2- For the predicted values, we calculate the sum of every column:

- Severity 1: 0 + 0 + 0 + 22654 = 22654 => 22654 predicted crashes
- Severity 2: 2 + 17 + 11003 + 0 = 11022 => 11022 predicted crashes
- Severity 2b: 559 predicted crashes
- Severity 3: 66 predicted crashes

A False Potisive (FP) is when the model predicts that an observation is of class A but it is actually not (Type I error).

A False Negative (NP) is when the model predicts that an obesrvation is not of class A but it actually is (Type II error).

Conclusions we can draw from the confusion matrix:

- Severity 1: all the predictions of Severity 1 were true.
- Severity 2: 17 + 2 = 19 FPs, i.e., predicted to be of Severity 2 but they are actually not. 17 are actually of Severity 2b and 2 are of Severity 3.
- Severity 2b: 1 FP and 17 FNs. The 1 FP was predicted to be of Severity 2b but it is of Severity 3, and the 17 FNs are the ones that were not predicted to be of Severity 2b (predicted to be of Severity 2), but they are actually of Severity 2b.
- Severity 3: 2 + 1 = 3 FNs. 3 are of Severity 3, but 2 were predicted to be of Severity 2, and 1 of Severity 2b.

### 4.1.2 Classification Report

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **1** | 1.00 | 1.00 | 1.00 | 22654 |
| **2** | 1.00 | 1.00 | 1.00 | 11003 |
| **2b** | 1.00 | 0.97 | 0.98 | 575 |
| **3** | 1.00 | 0.96 | 0.98 | 69 |

*Table 2: Logistic Regression Classification Report*

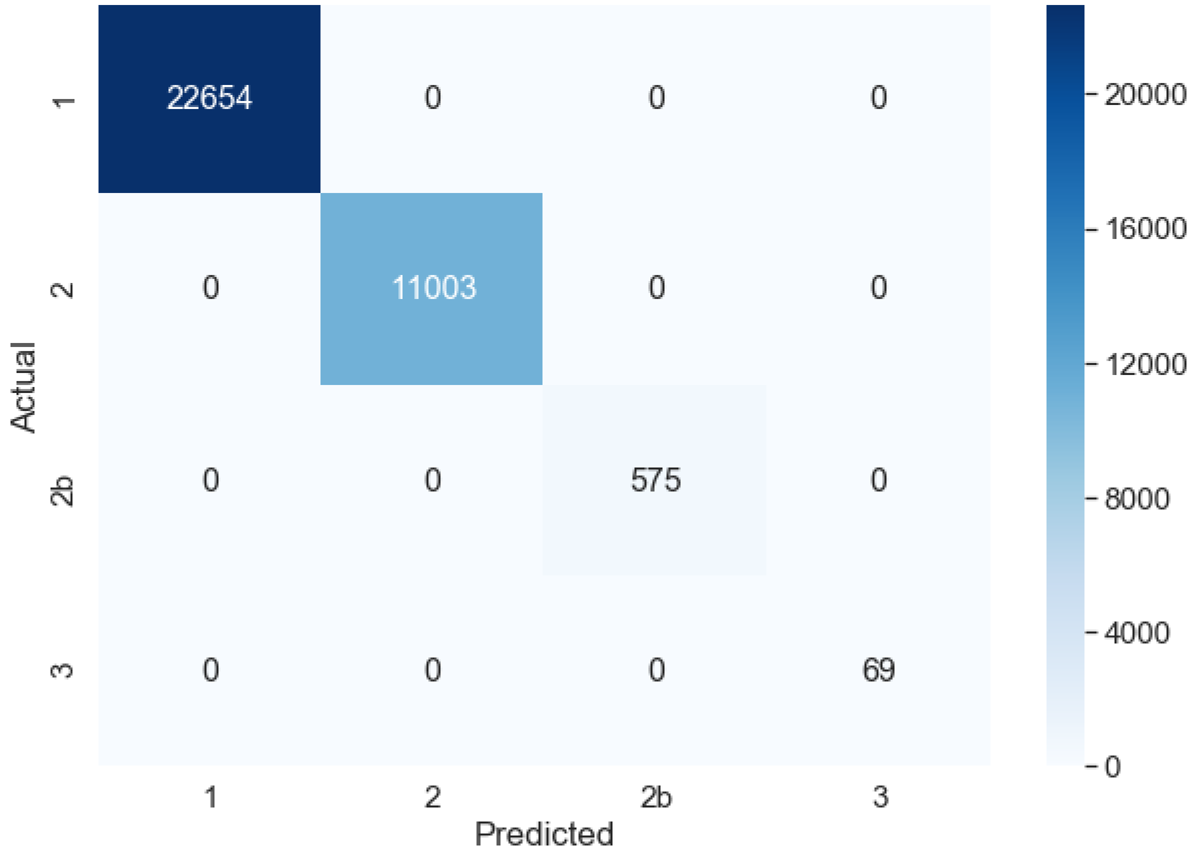
- Precision: measures the classifier’s exactness. Consider it as an answer to the assumption that for all instances classified as positives, what percent was correct?

- Recall: measures the completeness (the classifier’s ability to find all positive instances). For each class, it is given by TP / (TP + FN)
- F1-score: the best score is 1.0 and the worst in 0.0. The closer the score is to 1, the better the model is.
- Support: number of true instances for each class

Our Logistic Regression model had a 99.941% training accuracy and 99.958% testing accuracy.

## 4.2 Support Vector Machine

### 4.2.1 Confusion Matrix



| | 1 | 2 | 2b | 3 |
|---|---|---|---|---|
| 1 | 22654 | 0 | 0 | 0 |
| 2 | 0 | 11003 | 0 | 0 |
| 2b | 0 | 0 | 575 | 0 |
| 3 | 0 | 0 | 0 | 69 |

*Figure 10: SVM Confusion Matrix*

There are no FPs nor FNs for any label. Hence, this model has a 100% prediction accuracy over the testing set. We also calculated the training accuracy during our work and it was 100%.

### 4.2.2 Classification Report

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **1** | 1.00 | 1.00 | 1.00 | 22654 |
| **2** | 1.00 | 1.00 | 1.00 | 11003 |
| **2b** | 1.00 | 1.00 | 1.00 | 575 |
| **3** | 1.00 | 1.00 | 1.00 | 69 |

*Table 3: SVM Classification Report*

## 4.3 Decision Tree

### 4.3.1 Confusion Matrix

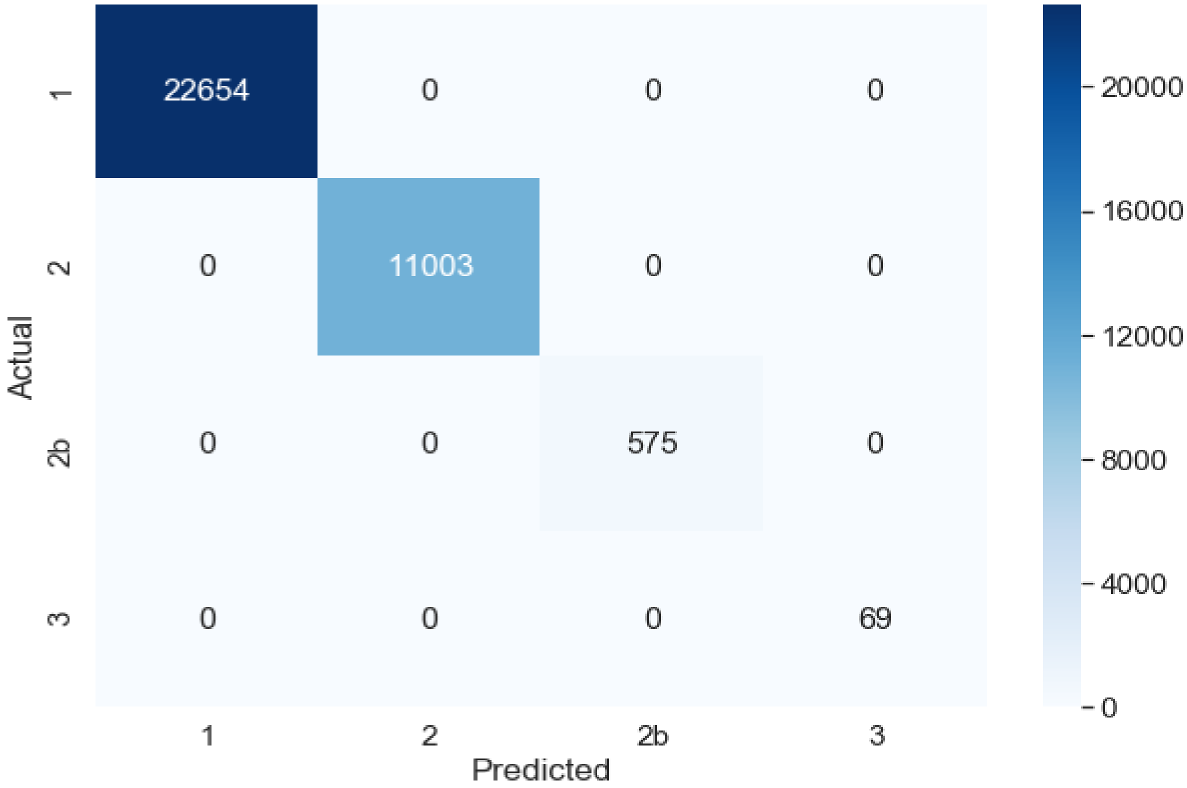


*Figure 11: Decision Tree Confusion Matrix*

### 4.3.2 Classification Report

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **1** | 1.00 | 1.00 | 1.00 | 22654 |
| **2** | 1.00 | 1.00 | 1.00 | 11003 |
| **2b** | 1.00 | 1.00 | 1.00 | 575 |
| **3** | 1.00 | 1.00 | 1.00 | 69 |

*Table 4: Decision Tree Classification Report*

100% training, as well as testing, accuracies.

## 4.4 Random Forest

### 4.4.1 Confusion Matrix

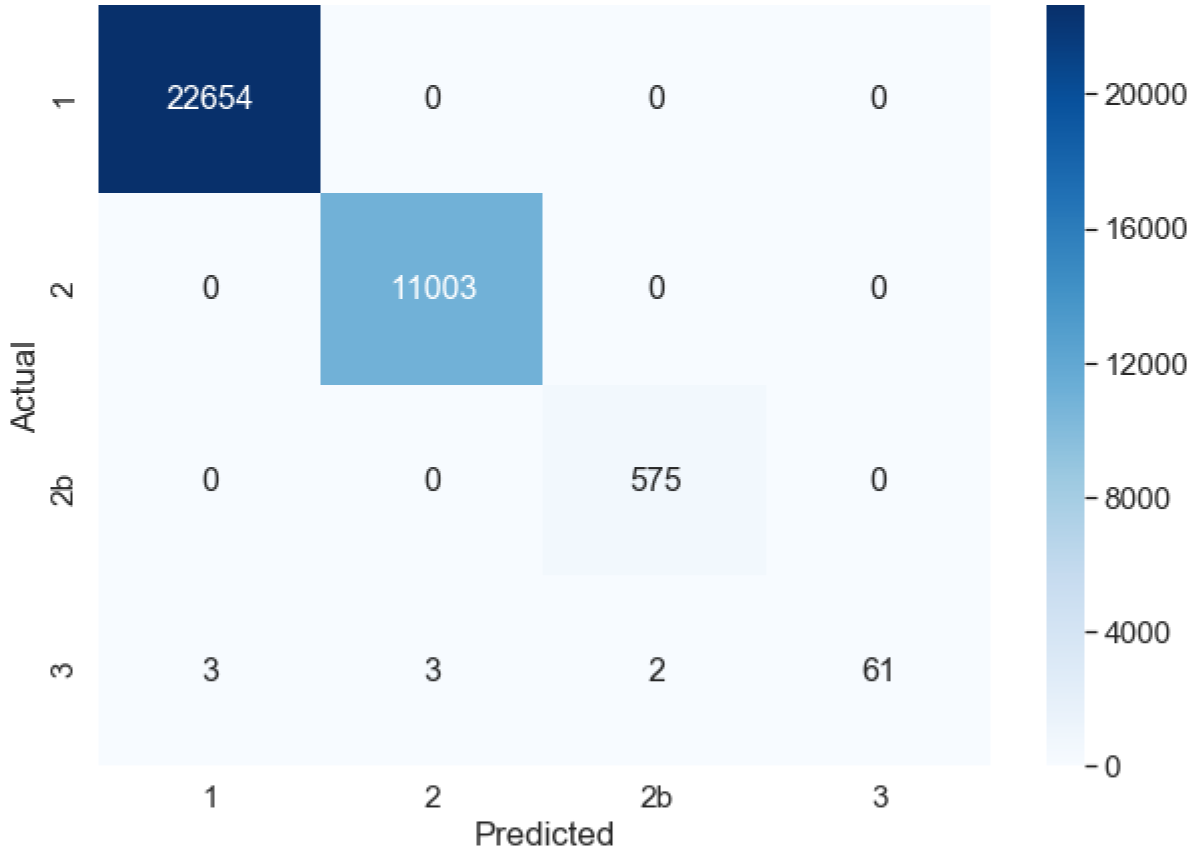


*Figure 12: Random Forest Confusion Matrix*

Predicted 1, 2, and 2b have 3, 3, and 2 FPs, respectively: every number of FPs was thought to be to the corresponding class (1, 2, and 2b) but they actually corresponds to class 3.

Predicted 3 has 8 FNs (8 that are actually of Severity 3 but predicted otherwise): 3 were predicted to be of Severity 1, 3 of Severity 2, and 2 of Severity 2b.

### 4.4.2 Classification Report

| | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| **1** | 1.00 | 1.00 | 1.00 | 22654 |
| **2** | 1.00 | 1.00 | 1.00 | 11003 |
| **2b** | 1.00 | 1.00 | 1.00 | 575 |
| **3** | 1.00 | 0.88 | 0.94 | 69 |

*Table 5: Random Forest Classification Report*

100% training accuracy and 99.976% testing accuracy.

# Chapter 5: Discussion

All of the 4 algorithms had an excellent accuracy score (above 99%) on both training and testing datasets. This means that our developed models are able to accurately predict an accident's severity. The above horizontal bar plot represents the accuracy score of each model in ascending order (on the graph, from the bottom to the top).

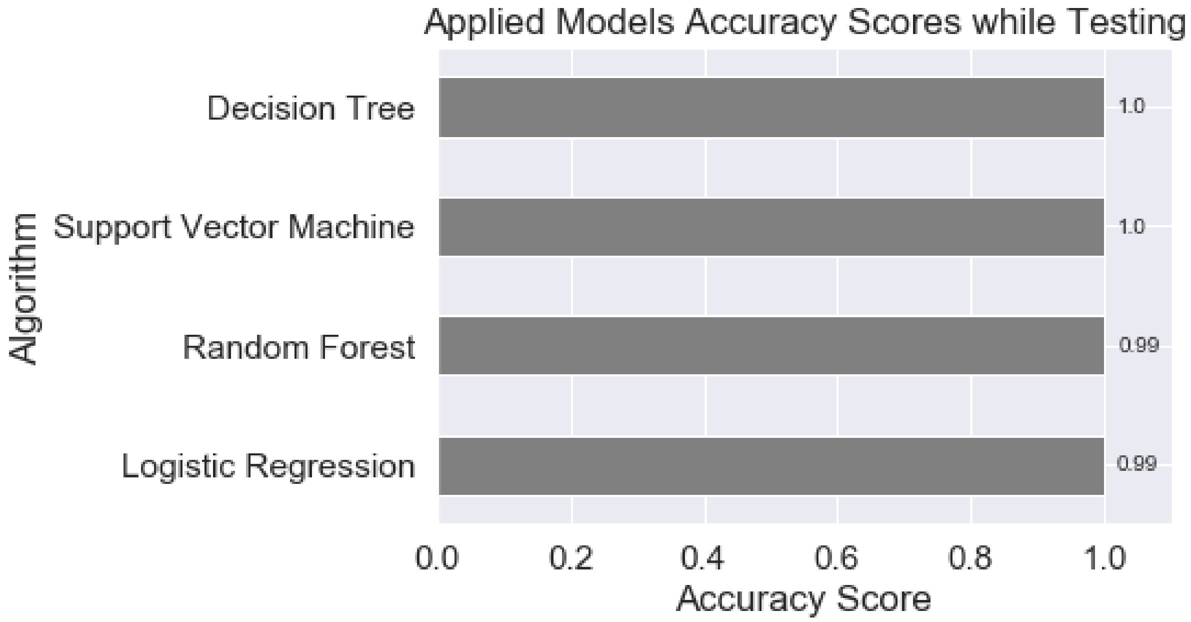


*Figure 13: Applied Models Accuracy Scores while Testing*

# Chapter 6: Conclusion

After preprocessing the data, we were able to test 4 algorithms, all of which showed excellent accuracy. The lowest accuracy achieved, on both training and testing datasets, exceeded 99%.

Our models were well trained and fit to the training data, and that it had excellent performance on the testing set. We could claim that our data was handled quite well.

Therefore, we can state that all of the four models can accurately predict a car accident's severity in the city of Seattle.