Predicting the severity of an accident — IBM Applied Data Science Capstone

Paul Moreira

October 2020

Contents

1 Introduction **3**

1.1 Background . . . 3

1.2 Problem . . . 3

1.3 Interest . . . 3

2 Data **3**

2.1 Data acquisition . . . 3

2.2 Data cleaning . . . 4

3 Exploratory Data Analysis **5**

3.1 Junction Type vs Severity . . . 5

3.2 Address Type vs Severity . . . 5

3.3 Collision Type vs Severity . . . 6

3.4 Under influence of drugs or alcohol vs Severity . . . 6

3.5 Weather vs Severity . . . 7

3.6 Road condition vs Severity . . . 7

3.7 Light condition vs Severity . . . 8

3.8 Speeding vs Severity . . . 8

4 Predictive modeling **9**

4.1 K-nearest neighbor . . . 9

4.2 Decision Tree . . . 10

4.3 Logistic Regression . . . 10

4.4 Performance of classification models . . . 11

5 Conclusion **12**

6 Future directions **12**

List of Figures

1 Junction type vs Severity . . . 5

2 Address type vs Severity . . . 6

3 Collision type vs Severity . . . 6

4 Under influence vs Severity . . . 7

5 Weather vs Severity . . . 7

6 Road condition vs Severity . . . 8

7 Light condition vs Severity . . . 8

8 Light condition vs Severity . . . 9

9 K value . . . 9

10 K-nearest neighbor confusion matrix . . . 10

11 K value . . . 10

12 Decision tree confusion matrix 11
13 Logistic Regression confusion matrix 11

1 Introduction

1.1 Background

According to the World Health Organization approximately 1.35 million people die each year as a result of road traffic crashes, and it costs most countries 3% of their gross domestic product. Traffic accidents cause economic losses to people and their families, and not only material losses because in the event of death or injuries, they affect several other aspects of the community. Starting from the premise that traffic accidents can be prevented, it is imperative to know the main factors that can cause an accident in order to develop strategies for the government or some organizations to act.
We will analyze a data set consisting on all type of collisions from 2004 to late 2020 in Seattle, provided by SPD and recorded by Traffic Records. By applying some machine learning techniques, we will predict the possible outcome of a traffic accident in terms of fatality.

1.2 Problem

In Seattle, as it is for most cities in the world, it is necessary to implement initiatives to reduce the rates and severity of traffic accidents, so this project aims to predict whether the outcome of an incident is severe or not based on available data.

1.3 Interest

Drivers, pedestrians, cyclists could benefit from the results of the prediction, because they could act under certain conditions that are very likely to cause an accident.
Government, traffic departments, police who need to take steps to create safe roads, build safer infrastructure, improve post-accident care for victims and raise awareness.

2 Data

2.1 Data acquisition

A comprehensive dataset consisting of road incident records that occurred between 2004 and October 16, 2020 in Seattle, obtained from the Seattle Open Data Portal 1. This dataset contains 221,524 rows and 40 columns, which describe the details of each incident, including location, severity, type of collision, fatality, date and time, weather, and road conditions, among others.

2.2 Data cleaning

Once the data set was downloaded, the analysis of the attribute information of the dataset was performed determining that, in the original form, the dataset is not suitable for data analysis. There are several issues with the dataset and it will be handled as follows:

Relevant features: The 9 most relevant features were selected so the rest of the characteristics were discarded. The kept features are:

- SEVERITYCODE: A code that corresponds to the severity of the collision.
- COLLISIONTYPE: Collision type.
- JUNCTIONTYPE: Category of junction at which collision took place.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N).
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol.
- WEATHER: A description of the weather conditions during the time of the collision.
- ROADCOND: The condition of the road during the collision.
- LIGHTCOND: The light conditions during the collision.
- SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)
- ADDRTYPE: Collision address type.

Incompleteness: There are missing values for most of the features even in the label that we want to predict, therefore, in order to build a model and determine how the features interact with the accident severity variable, we determined that it is necessary to remove the missing data entries of more than 2 key variables.
Missing values: We identified the missing values for each characteristic and analyzed them individually, for most of the characteristics it was necessary to replace the missing values with the highest frequency. For some features, the dataset contains only the positive values, so for the rest it was replaced by the value 'N'.
Correct data format: We proceed to convert data types to the proper format, as it is the case with integer features.
Normalized: For the some characteristics, it was necessary to normalize

the values unifying 'N' with 0 and Y with 'S'. Also when found overlapping values such as 'Other' were unified with the 'Unknown' value. To simplify the analysis, it was necessary to unify values in some characteristics based on the similarities and frequency.

3 Exploratory Data Analysis

First of all, it is important to know the data in order to observe the contribution that each characteristic has in the general dataset. In addition, through graphs we can contrast each of the characteristics with the target variable.

3.1 Junction Type vs Severity

The Figure 1 shows that when an incident is intersection related the risk of injury is increased.

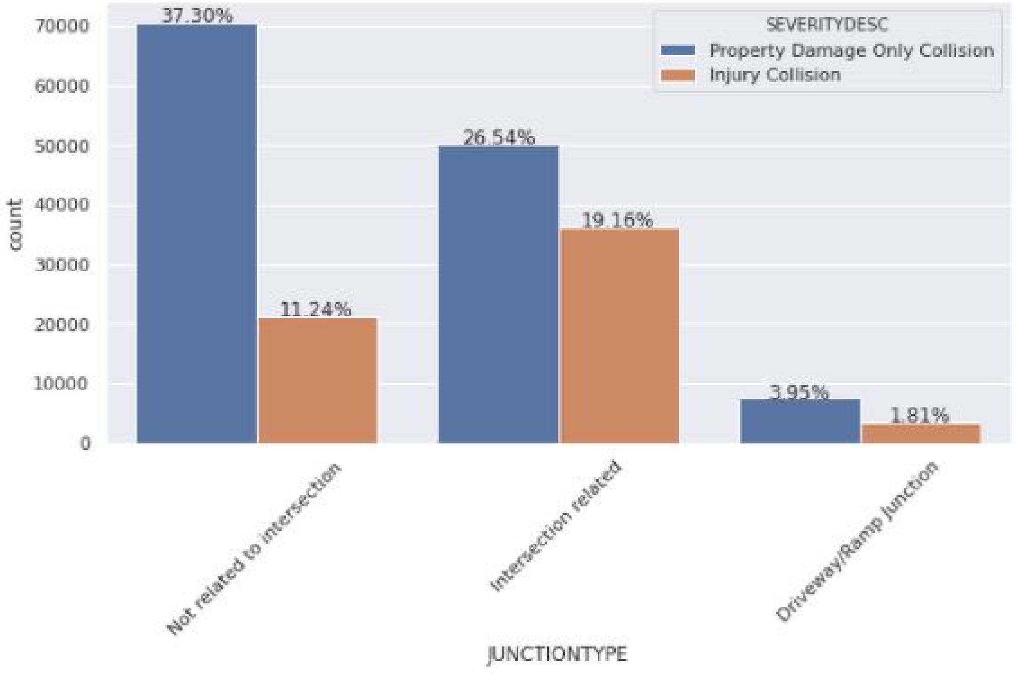

Figure 1: Junction type vs Severity

3.2 Address Type vs Severity

The Figure 2 confirms that when an incident is close to an intersection the risk of injury is increased.

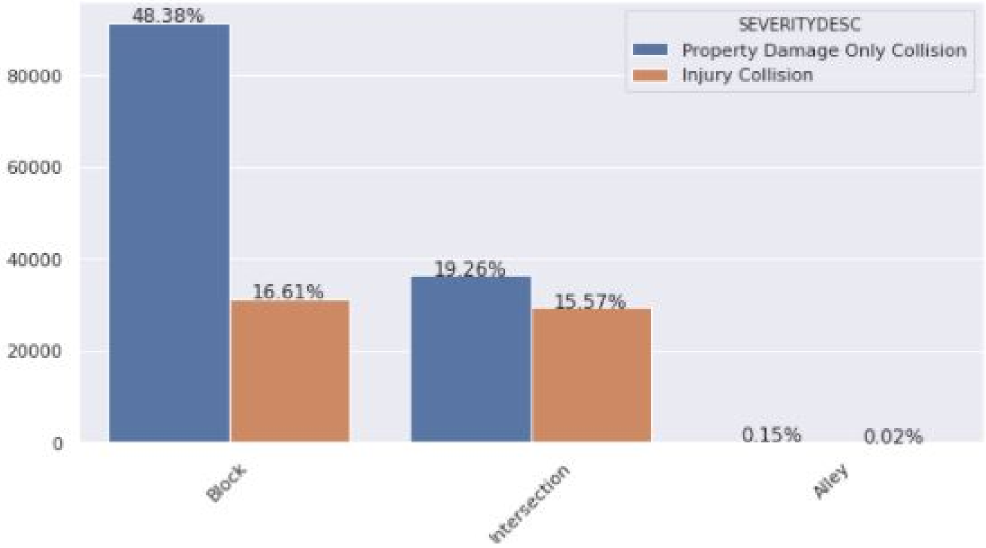

Figure 2: Address type vs Severity

3.3 Collision Type vs Severity

The Figure 3 shows that the risk of an injury is greater when the incident is against a cycle or a pedestrian.

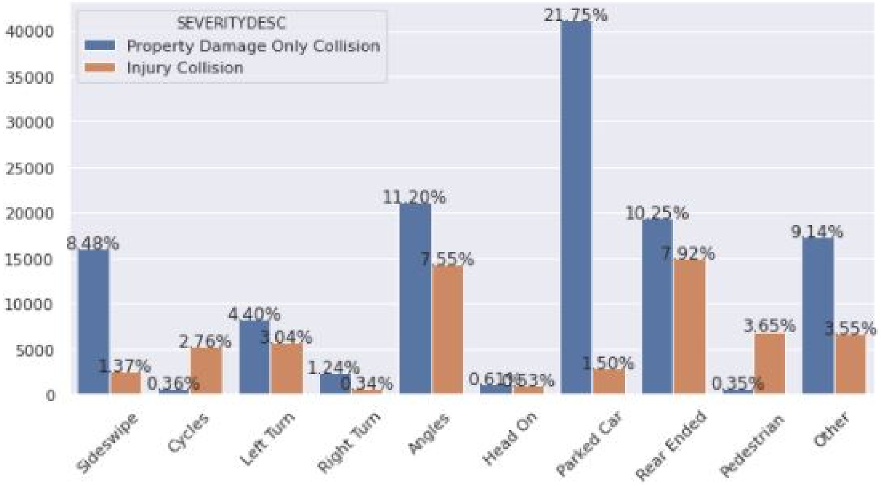

Figure 3: Collision type vs Severity

3.4 Under influence of drugs or alcohol vs Severity

Figure 4 shows that the risk of injury is high when the there are drugs or alcohol involved.

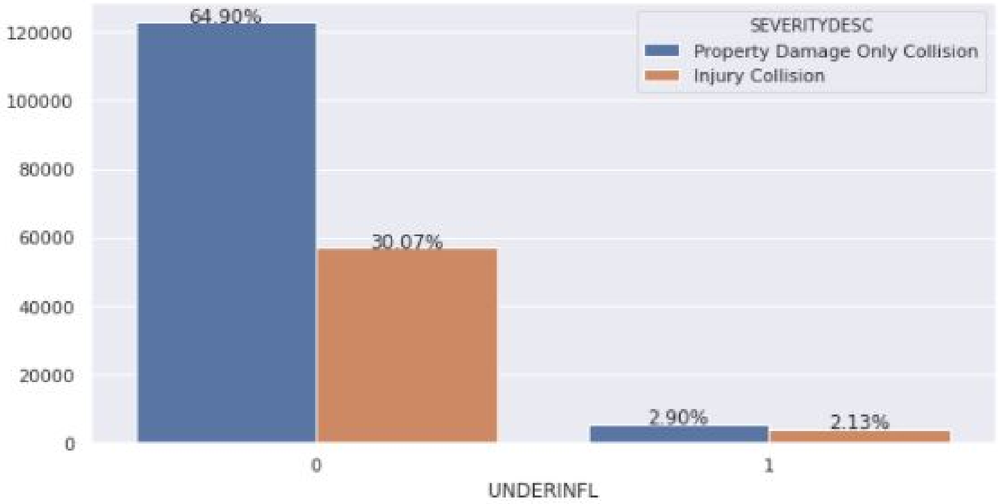

Figure 4: Under influence vs Severity

3.5 Weather vs Severity

Figure 5 does not show significant differences between the weather conditions and the severity of the accident..

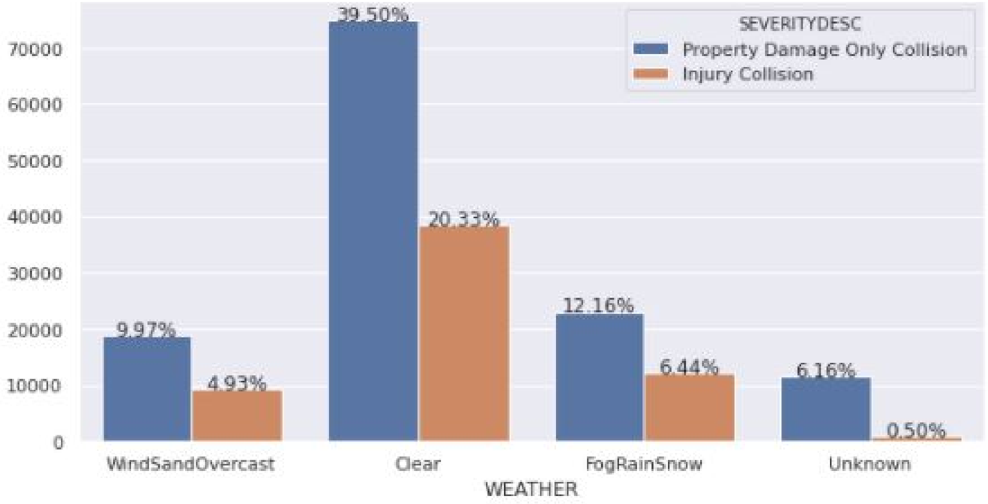

Figure 5: Weather vs Severity

3.6 Road condition vs Severity

Figure 6 also does not show significant differences between the road conditions and the severity of the accident.

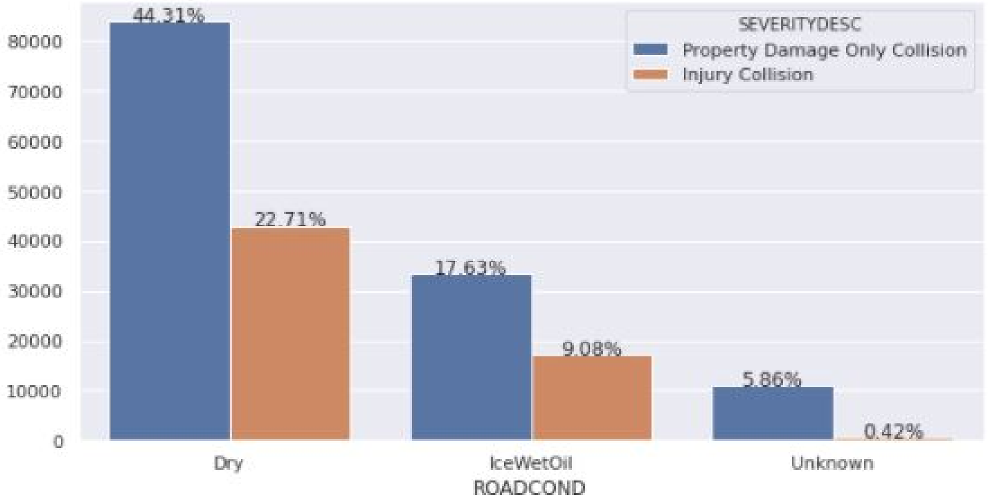

Figure 6: Road condition vs Severity

3.7 Light condition vs Severity

Figure 7 shows a slight increase in the severity of the incident when it is dark.

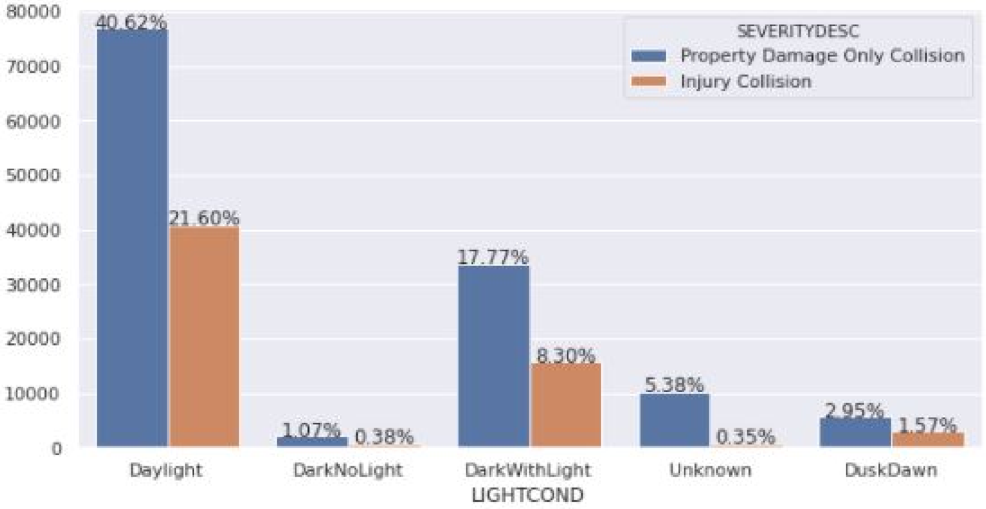

Figure 7: Light condition vs Severity

3.8 Speeding vs Severity

Figure 8 shows that when speeding is the cause of the accident, the risk of injury increases.

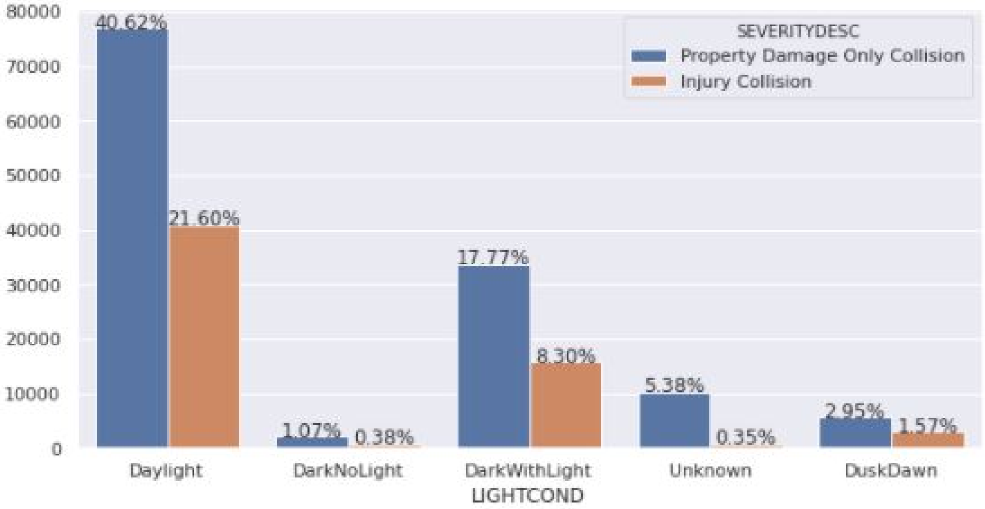

Figure 8: Light condition vs Severity

4 Predictive modeling

We will try to predict the classification of an accident fatality using the machine learning models K-nearest Neighbor, Decision Tree Analysis and Logistic Regression.

4.1 K-nearest neighbor

The accuracy of the k value increases as the value is also increase, therefore, the best k value is 9 with the accuracy of 0.63, as shown below.

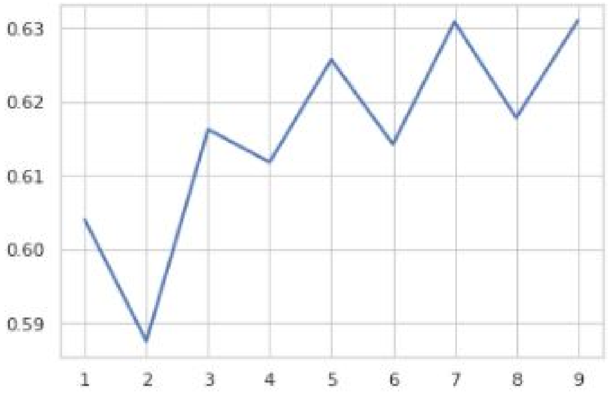

Figure 9: K value

The figure 10 shows the confusion matrix of the model using test data.

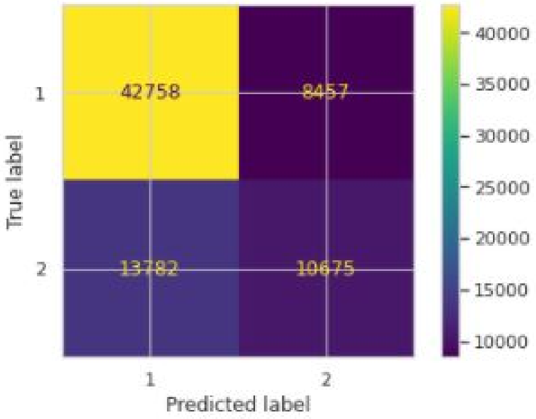

Figure 10: K-nearest neighbor confusion matrix

4.2 Decision Tree

The criterion chosen for the classifier was 'entropy' and the max depth was 24 with best accuracy of 70

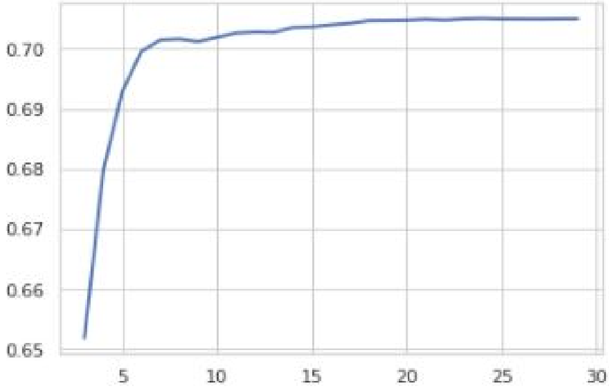

Figure 11: K value

The figure 10 shows the confusion matrix of the model using test data.

4.3 Logistic Regression

We use GridSearchCV to search the best parameters. The C used for regularization strength was '10.0' and penalty was "l2" with the accuracy of 70%, whereas the solver used was 'liblinear'.
The figure 10 shows the confusion matrix of the model using test data.

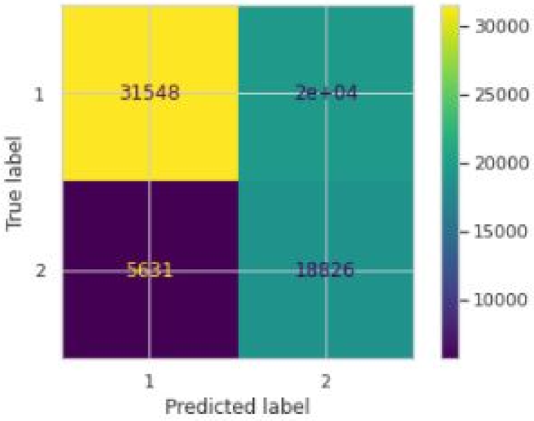

Figure 12: Decision tree confusion matrix

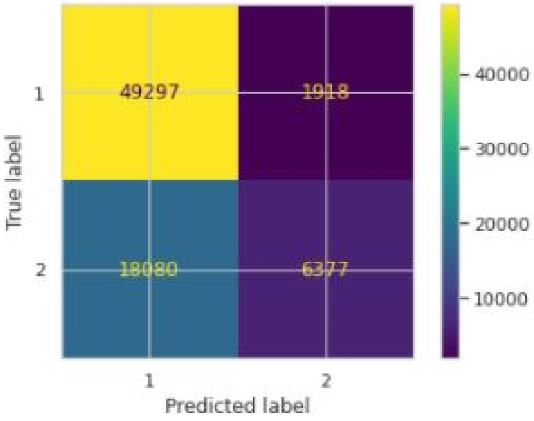

Figure 13: Logistic Regression confusion matrix

4.4 Performance of classification models

	K nearest neighbor	Decision tree	Logistic regression
Accuracy	0.65	0.72	0.70
F1 Score	0.63	0.72	0.74
True Positive	42758	31548	49297
True Negative	10675	18826	6377
False Positives	13782	5631	18080
False Negatives	8457	20000	1918

5 Conclusion

In this project we analyze the relationships between the severity of an accident and certain external characteristics such as weather, road and light conditions, collision and junction types among others. Through descriptive analysis it was possible to learn more about the features and data. Then, machine learning models were built to predict whether an accident has serious consequences or not. These models can be very useful in designing and implementing new politics and measures to avoid severities and fatalities in accidents.

6 Future directions

There were some characteristics that were not considered in this study, such as location and time characteristics, which may be important to consider in further studies. At various stages of the study, we found unbalanced data, so if we can improve the data set, for future performances the output from the models could get better. It is necessary to dig deeper to identify unrecognized factors and improve the accuracy of the models.