Predicting Car Accident Severity - A Case Study of Seattle, Washington D.C.

By,

Suhas V Londhe, 13th Sept 2020

Business Problem / Introduction

Road Accident refers to any accident involving at least one road vehicle, occurring on a road open to public circulation, and in which at least one person is injured or killed. Intentional acts (Murder, suicide) and natural disasters are excluded.
Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.
Analysing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These insights, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where some severe accidents can occur as well as saving both, time and money. In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.
Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.
The aim of this project is to predict the accident severity in Seattle, reasons for the accident, and where it could happen, in order to mitigate and limit future occurrences and ensure the safety of lives and properties.
Our targeted audiences are Seattle City Council, Government, and decision-makers, the general public, and Seattle Traffic Management Division-SDOT.

Key Facts

The US Department of Transportation recently reports that 871 Billion is the economic loss of motor vehicle crashes every year in the US.
This figure includes 271 Billion in economic loss and 594 Billion in harms from loss of lives and the pain.
Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.
Approximately 1.35 million people die each year as a result of road traffic crashes. The 2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020. Road traffic crashes cost most countries 3% of their gross domestic product. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.Socioeconomic status More than 90% of road traffic deaths occur in low- and middle-income countries. Road traffic injury death rates are highest in the African region. Even within high-income countries,

people from lower socioeconomic backgrounds are more likely to be involved in road traffic crashes.
Seattle is the largest seaport city on the West Coast of the United States. According to the data released in 2019, the metropolitan population stands at 3.98million. In July 2016, it was the fastest major growing city in the USA, with an annual growth rate of 3.1%.

General facts & the factors affecting the road accidents -

1. Age

Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

2. Sex

From a young age, males are more likely to be involved in road traffic crashes than females. About three quarters (73%) of all road traffic deaths occur among young males under the age of 25 years who are almost 3 times as likely to be killed in a road traffic crash as young females.

3. Alcohol consumption

Several studies ahows that drinking an alcohol & driving the vehicle leads to the accidents as alcohol directly impacts on the nerve cells of the brain; resulting in uncontrolled human behavior

4. Speeding

High speed driving is the another cause of road accidents. At times, high speed vehicle is unable to be controlled at the urgent situation, thus, results in a collision.
The evidence shows the risk of having a crash is increased both for vehicles traveling slower than the average speed, and for those traveling above the average speed.
The risk of being injured increases exponentially with speeds much faster than the median speed.
The severity / lethality of a crash depends on the vehicle speed change at impact.

5. Sleep deprivation

Various factors such as fatigue or sleep deprivation might increase the risk, or numbers of hours driving might increase the risk of an accident.

6. Road Design

The road or environmental factor was either noted as making a significant contribution to the circumstances of the crash, or did not allow room to recover. In these circumstances, it is frequently the driver who is blamed rather than the road; those reporting the collisions have a tendency to overlook the human factors involved, such as the subtleties of design and maintenance that a driver could fail to observe or inadequately compensate for.
In the UK, research has shown that investment in a safe road infrastructure program could yield a 1⁄3 reduction in road deaths, saving as much as £6 billion per year. A consortium of 13 major road safety stakeholders have formed the Campaign for Safe Road Design, which is calling on the UK Government to make safe road design a national transport priority.

7. Seat Belts

Research has shown that, across all collision types, it is less likely that seat belts were worn in collisions involving death or serious injury, rather than light injury; wearing a seat belt reduces the risk of death by about 45 percent.
Four driver behaviors (speed, stopping at intersections when the control light was amber, turning left in front of oncoming traffic, and gaps in following distance) were measured at various sites before and after the law. Changes in these behaviors in Newfoundland were similar to those in Nova Scotia, except that drivers in Newfoundland drove slower on expressways after the law, contrary to the risk compensation theory.

8. Maintenance

A well-designed and well-maintained vehicle, with good brakes, tires and well-adjusted suspension will be more controllable in an emergency and thus be better equipped to avoid collisions. Some mandatory vehicle inspection schemes include tests for some aspects of roadworthiness.
Common features designed to improve safety include thicker pillars, safety glass, interiors with no sharp edges, stronger bodies, other active or passive safety features, and smooth exteriors to reduce the consequences of an impact with pedestrians.
In the early 1970s, British Leyland started an intensive programme of vehicle safety research, producing a number of prototype experimental safety vehicles demonstrating various innovations for occupant and pedestrian protection such as air bags, anti-lock brakes, impact-absorbing side-panels, front and rear head restraints, run-flat tires, smooth and deformable front-ends, impact-absorbing bumpers, and retractable headlamps. Design has also been influenced by government legislation, such as the Euro NCAP impact test.

9. Drug use

Including some prescription drugs, over the counter drugs (notably antihistamines, opioids and muscarinic antagonists), and illegal drugs.

10. Use of mobile phones while driving

Drivers using mobile phones are approximately 4 times more likely to be involved in a crash than drivers not using a mobile phone. Using a phone while driving slows reaction times (notably braking reaction time, but also reaction to traffic signals), and makes it difficult to keep in the correct lane, and to keep the correct following distances. Hands-free phones are not much safer than hand-held phone sets, and texting considerably increases the risk of a crash.

11. Traffic Law enforcement Failure

Every country in the world have their own laws, rules & regulations on transportation; such as road, coastal, air etc. It is necessary to all the citizens that all the laws should be strictly followed in order to avoid any accident. Failure of traffic law enforcement will result in continual increase of road accidents rather than the decrease.

12. Immediate medical treatment failure to humans suffering from accident

It is observerd that, after the accident, it takes a lot of important time to give the emergency medical treatment to the humans surviving from the road accident. Some seriously injured people lose their life due to failure of ambulance, emergency treatment to the people.

Data

The goal of this section is to indicate the sources where the data has been collected from as well as describe the meaning of each feature in the acquired dataset.
The source of the data is Seattle Traffic Management Division-SDOT- (A US state Government Agency), as provided through a link on Coursera in CSV format. This particular dataset starts from 2004 and is updated weekly till this present time.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
The above link is used to download the dataset used in this project. It contains the data with the road collisions.
This data was provided by the Seattle Police Department and recorded by the Traffic Records Department. The dataset consists of 38 independent fields and 194673 records, which includes both numerical and categorical data. The dependent field or label for the data set is SEVERITY CODE, which describes the fatality of an accident.

Data Description

There are 38 features (attributes) in the dataset and 194673 records (rows) of car accidents. The first few columns provide several identification keys unique for each incident, this information is not usable,hence, it will be ignored in the analysis.

The location of the accident is described by three columns, two of which refer to the 'latitude' and 'longitude' values while the third one contains the address.

The information contained in the data includes - Speed, Collision Type, Weather situation, Severity Code, Location, Address Type, Road Condition, Status, Vehicle etc.

The date and time when the incident occurred is given by two columns, one containing only the date while the other containing the date and time of the collision.

The information is also available in the dataset on whether the accident occurred at an intersection, a block or an alley, the type of the junction and the unique keys of each intersection, crosswalk road segment where the accident occurred. There are several features showing variables such as the weather, the road condition, the lighting condition, whether the driver was inattentive, speeding was under the influence of alcohol or the pedestrian right of way was not granted.

It contains information about number of pedestrians, cyclists and vehicles involved in the accident and whether there was a collision with parked cars.

State Collision Code which uniquely describes each type of collision using a numeric value is present in a dataset.

Severity of each accident is given as a binary variable with the value '1' if the accident was less serious and '2' if the accident was severe. "Severity Code" column will be used from this dataset as a dependent variable / target variable that will be predicted in this project.

Feature Selection

The goal of this study is the development of a model that predicts the severity of an accident using realtime data, this puts some limits on the type of features that can be used since not all of the dataset’s columns contain information that can be obtained in real-time.

The dependent variable or target variable for the data set is "SEVERITY CODE", which describes the fatality of an accident.

Other attributes that have been taken into consideration for the analysis are as follows –

1. LOCATION - Description of the general location of the collision,
2. ADDRTYPE - Collision address types are 'Alley', 'Block' or 'Intersection'
3. ROADCOND - The condition of the road during the collision (for example, 'Wet', 'Dry', 'Snow/Slush', 'Ice', 'Sand/Mud/Dirt', 'Standing Water', 'Oil',etc.),
4. WEATHER - A description of the weather conditions during the time of the collision, (for example, 'Overcast', 'Raining', 'Clear', 'Snowing', 'Fog/Smog/Smoke', etc.),
5. JUNCTIONTYPE - Category of junction at which collision took place (for example, 'At Intersection (intersection related)', 'Mid-Block (not related to intersection)',etc.),
6. PERSONCOUNT - The total number of people involved in the collision (reflected as an integer),
7. VEHCOUNT - The number of vehicles involved in the collision (reflected as an integer),
8. LIGHTCOND - The light conditions during the collision (for example, 'Daylight', 'Dark - Street Lights On', 'Dark - No Street Lights', 'Dusk', 'Dawn', etc.)
9. SPEEDING - Whether or not speeding was a factor in the collision.

Data Preparation

The process of data preparation includes removing or filling missing values, converting features into certain formats and making sure the data is balanced. This step is required in most algorithms in order for them to be implemented correctly and can significantly boost their accuracy.

Data with status-unmatched is removed from the dataset in order to avoid unnecessary data.

Categorical features are converted to numerical values.

Data with missing feature values are removed.

Data imbalance is removed from the dataset in order to reduce the error in the model & increase the accuracy for the correct predictions

The dataset is slightly imbalanced, with 136,485 data points with Severity Code 1 and only 58,188 data points with Severity Code 2. Therefore, methodology of upsampling is used to remove data imbalance to end up with an equal number of data points with Severity Codes 1 and 2 (i.e 136,485 data points each).

Python libraries such as 'Pandas', 'Numpy' are used for the data preparation.

Methodology

1. Required Python libraries are imported –

Lets import the Python libraries required for this project

```
import pandas as pd
import numpy as np
from sklearn import metrics
from sklearn.metrics import classification_report, confusion_matrix
import matplotlib.pyplot as plt
from matplotlib.ticker import NullFormatter
import matplotlib.ticker as ticker
from sklearn import preprocessing
%matplotlib inline
from sklearn.utils import resample
from sklearn.ensemble import ExtraTreesClassifier
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.feature_selection import SelectFromModel
from sklearn.metrics import accuracy_score
import seaborn as sns
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_curve, auc
import matplotlib.image as mpimg
from sklearn import tree
from sklearn.tree import export_graphviz
from sklearn import svm
from sklearn.ensemble import RandomForestClassifier
import matplotlib as mpl
```

2. ‘CSV’ file is imported & converted into the dataframe as -

Load the data from the CSV file using 'Pandas' library

```
df = pd.read_csv("E:/Data Science class/IBM Data Science/My Project/Data-Collisions.csv")
df.head()
```

```
C:\Users\Lenovo\Anaconda3\lib\site-packages\IPython\core\interactiveshell.py:2728: DtypeWarning: Columns (33) have mixed types.
Specify dtype option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDF
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	

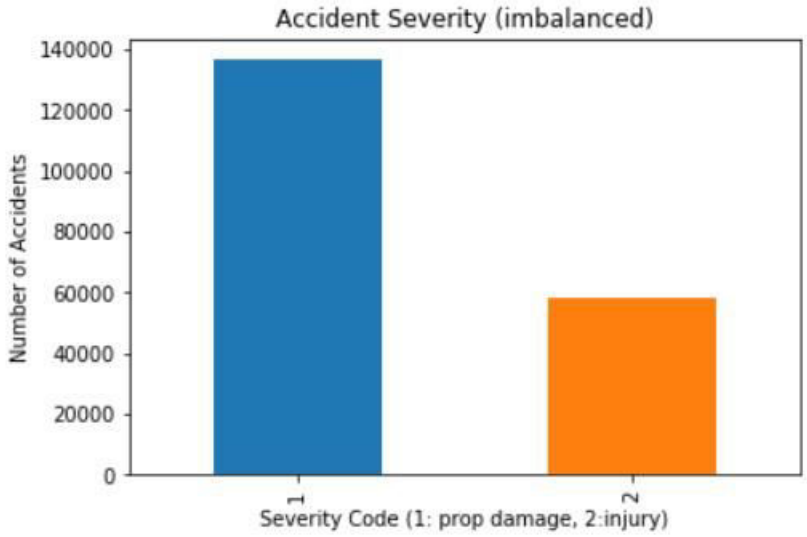
5 rows × 38 columns

3. Whether the data is balanced or imbalanced is checked.

Now, we will check the data imbalance

```
df.SEVERITYCODE.value_counts().plot(kind='bar')
plt.xlabel('Severity Code (1: prop damage, 2:injury)')
plt.ylabel('Number of Accidents')
plt.title('Accident Severity (imbalanced)')
```

```
Text(0.5,1,'Accident Severity (imbalanced)')
```



From the above bar graph, we notice that ther is a data imbalance

4. Undersampling is done on the dataset in order to get rid of the oversampling. In this process, minority class data is upsample in order to make it in the same proportion.

In [169]:
```
# Oversampling
# Separate majority and minority classes
df_majority = df[df.SEVERITYCODE==1]
df_minority = df[df.SEVERITYCODE==2]

#  We will Upsample minority class in order to remove the oversampling of the data

df_minority_upsampled = resample(df_minority,
                                 replace=True,     # sample with replacement
                                 n_samples=136485,    # to match majority class
                                 random_state=123) # reproducible results

# Combine majority class with upsampled minority class

df_upsampled = pd.concat([df_majority, df_minority_upsampled])

# Display the new class counts

df_upsampled.SEVERITYCODE.value_counts()
```

Out[169]:
```
2    136485
1    136485
Name: SEVERITYCODE, dtype: int64
```

5. Missing value treatment – It is done by replacing ‘Nan’ values.

In [178]:
```
# Replacing NaN value with the keyword "Unknown"
df2['WEATHER'].replace(np.NaN, "Unknown", inplace=True)
```

In [179]:
```
# Replacing Unknown and Other by Clear, the most frequent value of the column
encoding_WEATHER = {"WEATHER":
                                {"Clear": 1,
                                 "Unknown": 1,
                                 "Other": 1,
                                 "Raining": 2,
                                 "Overcast": 3,
                                 "Snowing": 4,
                                 "Fog/Smog/Smoke": 5,
                                 "Sleet/Hail/Freezing Rain": 6,
                                 "Blowing Sand/Dirt": 7,
                                 "Severe Crosswind": 8,
                                 "Partly Cloudy": 9}}
df2.replace(encoding_WEATHER, inplace=True)
df2['WEATHER'].value_counts()
```

Out[179]:
```
1    132139
2     33145
3     27714
4       907
5       569
6       113
7        56
8        25
9         5
Name: WEATHER, dtype: int64
```

After the features are selected, they are taken in a final dataset for an explanatory data analysis to figure out more about their effects. The focus is on identifying the feature conditions that have a bigger effect on the severity that leads to injuries. To do so, the dataset is filtered further and the corresponding values of features are sorted.

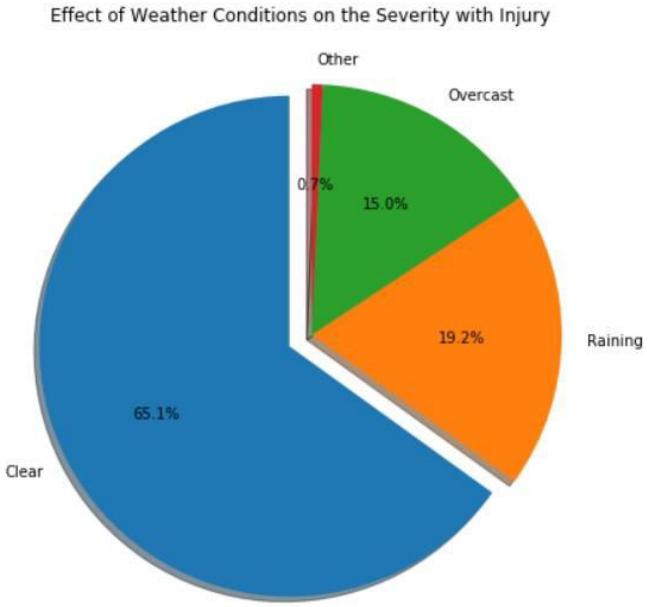
In the next step, the features are processed for predictive modeling analysis. 4 machine learning models are created using the classification techniques as listed below:

Decision Tree Logistic Regression Random Forest K-Nearest Neighbors (KNN) The created models are tested and then evaluated based on their accuracy score to find the best accurate model.

EDA (Exploratory Data Analysis)

Relationships between different independent variables with dependant variable is visualized using pie chart.

In [214]:
```
labels = 'Clear', 'Raining', 'Overcast', 'Other'
sizes = [37856, 11176, 8745, sum(Sev_2_w[3:9])]
explode = (0.1,0, 0, 0)
fig1, ax1 = plt.subplots(figsize=(15,7))
ax1.pie(sizes, explode=explode, labels=labels, autopct='%1.1f%%',shadow=True, startangle=90)
ax1.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.title('Effect of Weather Conditions on the Severity with Injury', y=1.05)
plt.show()
```



Modeling, testing & evaluation

Train-Test Split

Dataset is divided into training & testing set. 80% data is used for training the model & 20% data is used for testing the model using train-test split method.

Train-test split

In [234]:
```
#Splitting the data into train-test sets
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (155738, 10) (155738,)
Test set: (38935, 10) (38935,)
```

- **4 machine learning algorithms are used in modelling. We will only see decision tree modelling in this report.**

Decision Tree

In [240]:
```
#Model Building using decision tree classifier
DTree = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
DTree.fit(X_train,y_train)
```

Out[240]:
```
DecisionTreeClassifier(criterion='entropy', max_depth=4)
```

In [241]:
```
#Prediction of the model
yhat = DTree.predict(X_test)
yhat
```

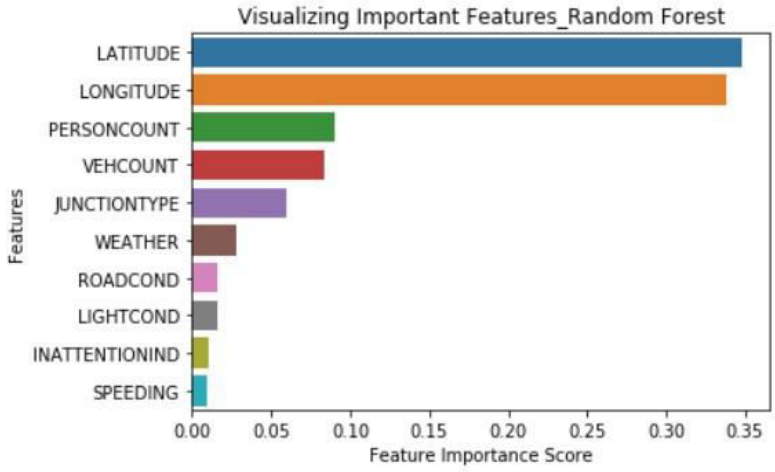
Out[241]:
```
array([1, 1, 1, ..., 1, 1, 1], dtype=int64)
```

Evaluation - Decision Tree

In [242]:
```
#Evalaution of the model
print("Decision Trees's Accuracy is: ", metrics.accuracy_score(y_test, yhat))
```

```
Decision Trees's Accuracy is:  0.7429562090663927
```

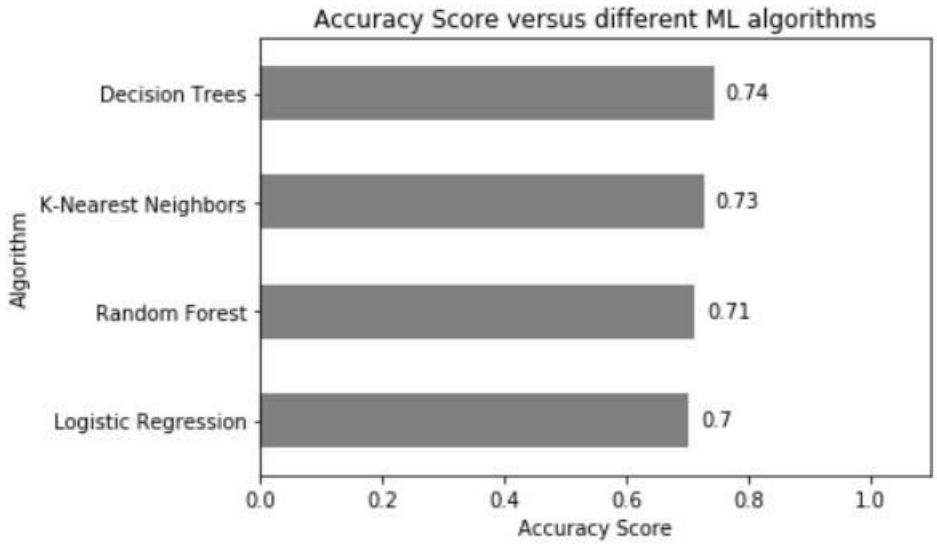
- **We will visualize the important independent features involved in modelling with random forest as –**



From the above graph, it is clear that 'LATITUDE' is the most important predictor & 'SPEEDING' is the least.

Results & discussion

We plotted the accuracy score for all machine learning algorithms. We found that, Decision Tree algorithm gave the highest accuracy amongst all 4 algorithms.



Conclusion

After comparing the score of accuracies obtained by the different machine learning algorithms K-Nearest Neighbors, Decision Tree, Logistic Regression, and Random Forest; Decision Tree algorithm has been proved to give the better accuracy.

During the modeling with K-Nearest Neighbors classifier, it was observed that the computer required much more time. But it took less time to execute the decision tree modeling. This can also represent better effectiveness and compatibility of the decision tree for handling this given dataset.

In this study, supervised machine learning is applied to predict car accident severity. The imbalanced dataset is initially balanced, and the raw data is analyzed and prepared in different steps to be fed into the machine learning models. In parallel, an explanatory data analysis is done to gain more insights into the relationship between the features and the severity of the accidents.

Four machine learning algorithms (K-Nearest Neighbors, Decision Trees, Logistic Regression, and Random Forest) are applied in which the decision tree has shown better compatibility with the dataset, resulting in higher accuracy (0.74).

One idea for future work can be applying feature selection algorithms such as LASSO to better select features. Enriching data further with other approaches to handling missing values can potentially improve prediction accuracy. Furthermore, testing other attributes related to car drivers, such as their age, can also be useful in better predicting car accident severity. Last but not least, performing k-fold cross-validation to obtain the optimal value of K in the KNN classifier may increase the accuracy as well.