CAR ACCIDENT SEVERITY

Final project

1. Introduction / Business problem

1a. Report

Road Traffic Injuries (RTIs) are a major public health problem. The World Health Organization (WHO) reports that the number of deaths due to road accidents has exceeded one million in recent years. (1)

The problem is huge and concerns everyone, a fact that makes the analysis of road accidents necessary. Forecasts have been rising in recent decades. (2) The analytical approach to the data will provide answers that can be used to identify and predict factors that affect the severity of road accidents.

The most popular scientific method is Machine Learning, because it has the ability to identify existing patterns in data and to predict, through the creation and evaluation of different algorithms.

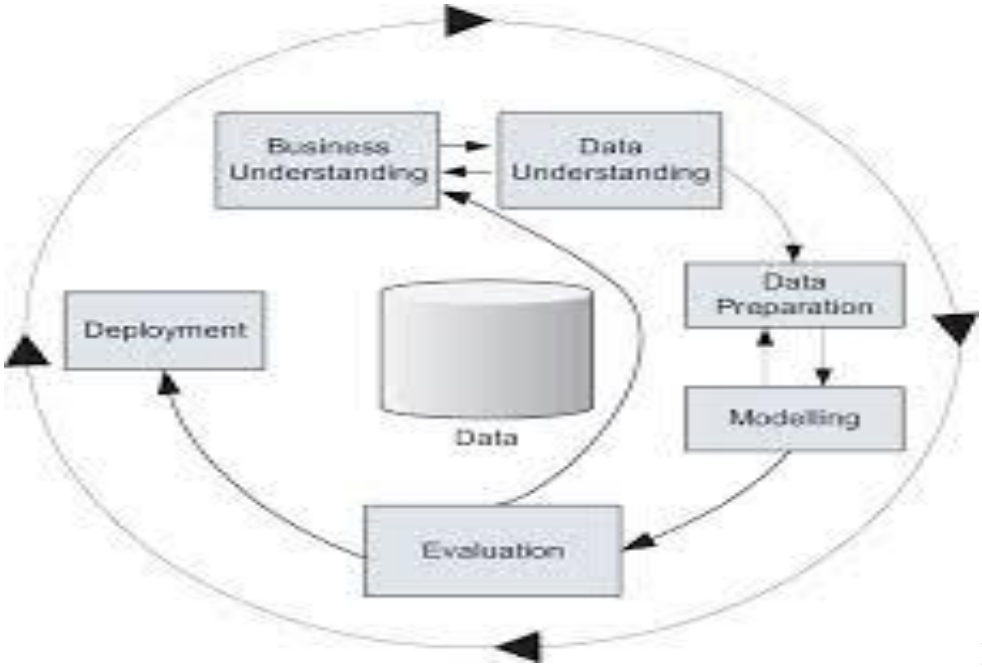
Furthermore, Machine Learning can manage another problem that arises, the large amounts of data that are generated, since road accidents are frequent and increasing at a rapid rate.

1b. Expectancy

This project aims to predict whether an accident that happens under a specific set of circumstances will be an accident limited to *property damage* or if it will include some form of *physical injury* to the driver and/or the passengers.

The aim is therefore to predict the severity of the accident through training and assessment machine learning algorithms, with the help of a data set, including recording provided by SDOT Traffic Management Division, Traffic Records Group in Seattle, United States.

As an approach to achieve this goal, the interprofessional standardized data mining process (CRISP-DM) (Figure 1) will be applied. The data will therefore be well understood and prepared before being fed for the forecast modeling analysis in the next steps.



F1, (3)

1c. Involved

The analysis is addressed to those involved in road traffic such as:

- National Emergency Center,
- Public Health Department
- Traffic Control Police authorities
- Infrastructural Development & Management Authorities
- Traffic Control Police Authorities,
- Roadside assistance services,
- Insurance Companies,
- Taxpayers & Travelers

in order to guide the stakeholders to decrease the *property damage* or/and *physical injury* by improving safety margins and ways to deal with the severity of the accident.

References:

(1) World Health Organization: http://www.who.int

(2) Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights, S. Moosavi et al., SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

(3) Cross-industry Standard Process for Data Mining (CRISP-DM), https://www.oreilly.com/library/

2. Data

2a. Origin of data

Based on the process (CRISP-DM), we can download data that will be utilized by SDOT Traffic Management Division, Traffic Records Group, Seattle, United States).

In the file (Data-Collisions.csv) (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) the column named "Severity Code" consists of two values:

- 1 = property damage
- 2 = injury

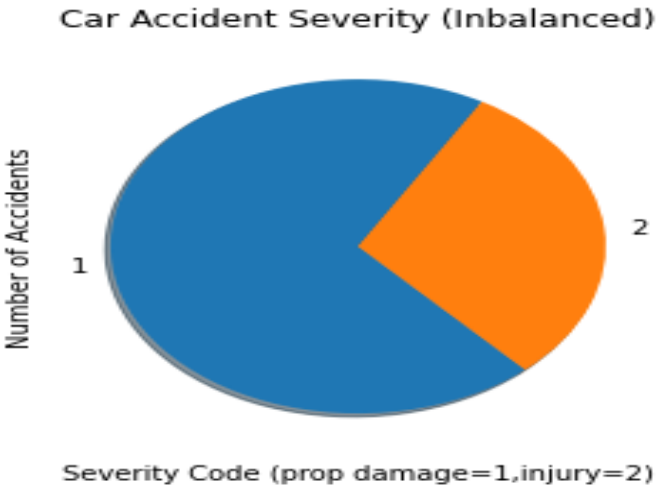
Also, other columns report various accident conditions such as: location, weather, light, road, types of collision, etc.

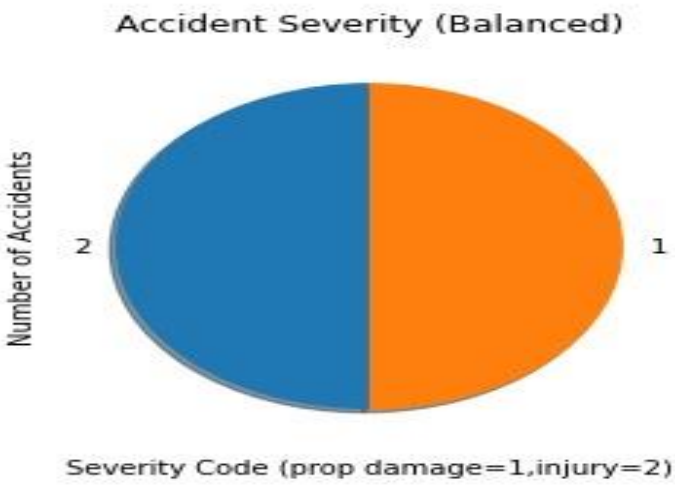
2b. Data balancing

We observe that the targeted variable "Severity Code" has more references to value 1 = property damage than to value 2 = injury.

So we have to balance the two values so as not to be led to wrong conclusions.

Sampling is a widely used technique to address this issue. It consists of removing random observations from the majority class to prevent the dominant signal of the learning algorithm (under sampling) or the accidental repetition of observations from the minority class to amplify the signal (over sampling). 2 Under sampling, a large amount of data will be lost, which can later be used to predict severity. Therefore, the hyper-sampling technique is preferred and applied in this project.





2c. Clearing the data

We notice that there are many empty "NaN" cells, which we will replace with the values that appear in the maximum range.

In addition, in order to develop regression models, certain variables must be converted to index variables.

Some other columns contain data values regardless of the analysis we want to perform or are overshadowed by other data, while some will be scrolled and renamed.

2d. Selected Independent / Variable Forecasts:

1. LONGITUDE
2. LATITUDE
3. INDIVIDUAL (total number of people involved in the conflict)
4. VEHCOUNT (total number of vehicles involved in on collision time)
5. JUNCTIONTYPE (category of the intersection at which the collision happened)
6. WARNING (whether or not the collision is due to carelessness)
7. WEATHER (weather conditions on collision time)
8. ROADCOND (condition of the road on collision time)
9. LIGHTTCOND (lighting conditions on collision time)
10. SPEED (whether speed was a collision factor or not)

3. Methodology

We need to draw our attention to those values that have the greatest impact on the severity of injuries.

For this reason the data is filtered and sorted according to the level of severity. We introduce the Folium package and pair a feature set and a pointer = circle so that the Seattle city map shows the accident sites. Pie charts are then used to display the numerical sizes. As mentioned in Chapter 2, the following are studied:

- Weather conditions
- The number of people involved in the accident
- The number of vehicles
- The types of road junctions
- The carelessness of the driver
- The effect of light
- The effect of road conditions
- The driving speed

All the results are collected in the section "Exploratory Data Analysis".

The results are then processed, normalized and classification techniques are applied where various machine learning algorithms are tested to create the model.

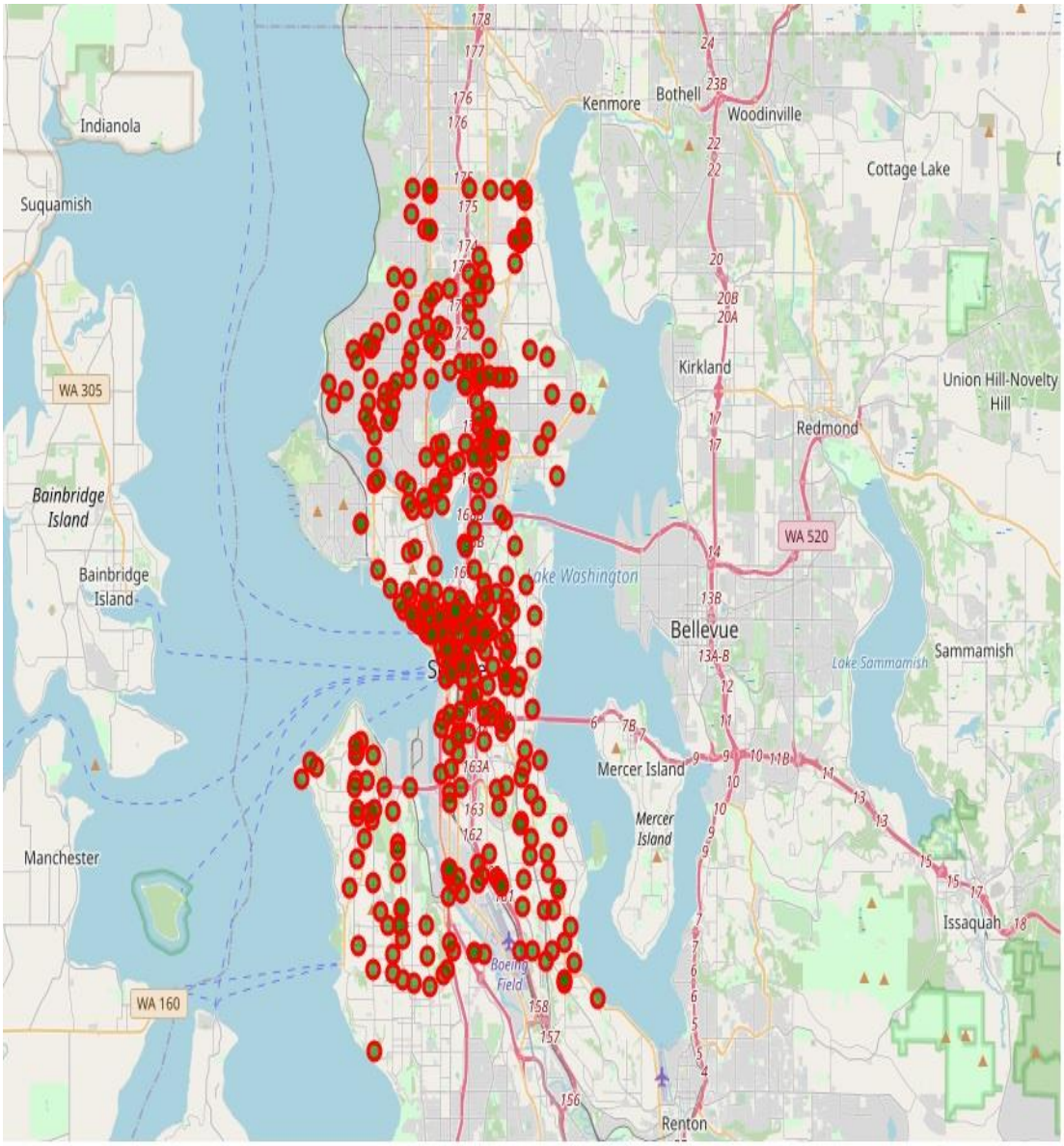
1. K-Nearest Neighbors (KNN) applies. The KNN operates on the basis of the classification of the nearest K points near the point to be predicted.
2. The Decision Tree is applied, in which each internal node is a representation of a test on a attribute, each branch is the representation of the test result and each leaf node is the representation of a class tag.
3. Accounting Regression is applied
4. Random Forest is applied to judge the interaction of different features

5. The Accuracy Measure is applied, which shows us the set of correct predictions.

The results of the Predictive Analysis are collected in the Modeling, Testing and Evaluation Section.

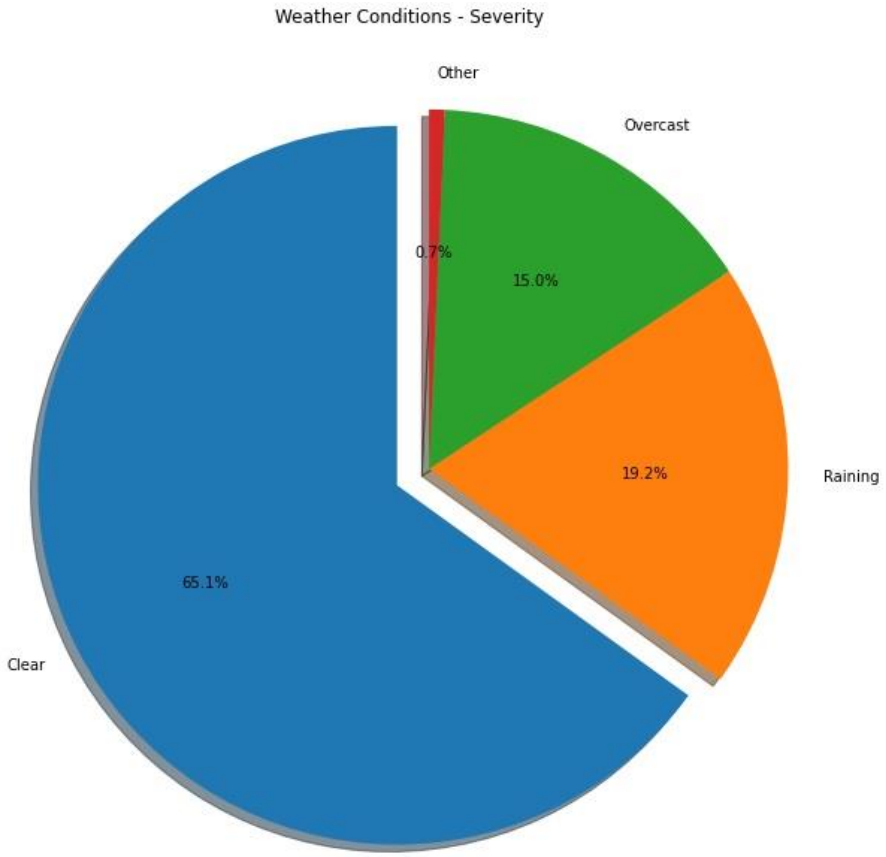
3.1. Exploratory Data Analysis

A map of Seattle Area with a limit of 400 accidents data points is attached.



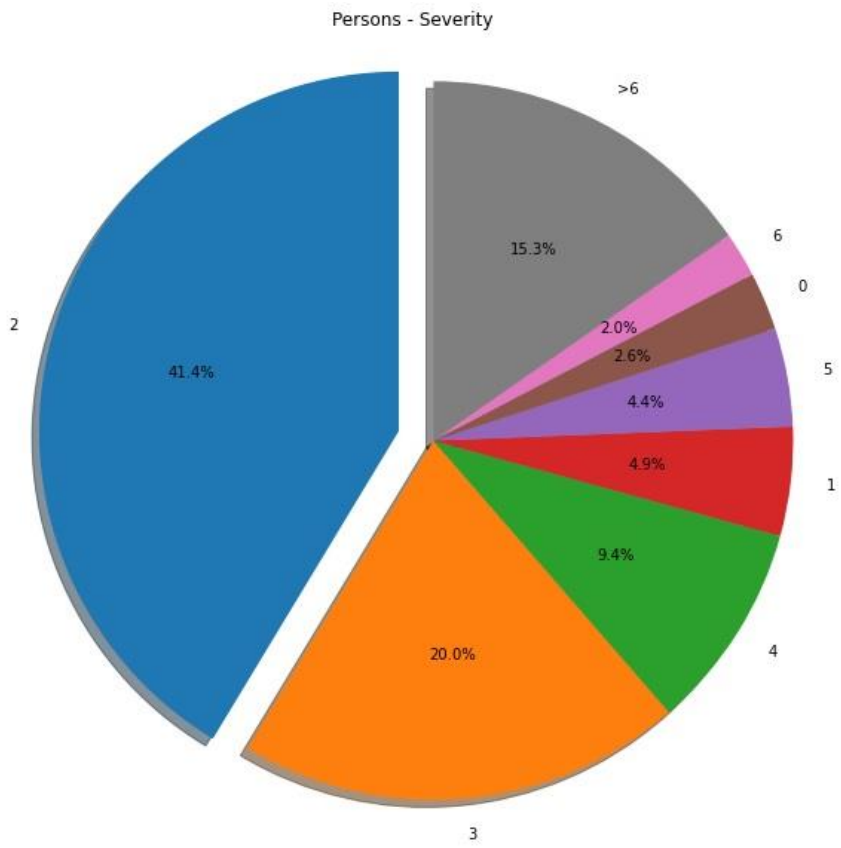
Weather Conditions – Severity

About 65% of serious injury accidents occurred in clear conditions in Seattle. Rainfall of about 19% and cloudy skies of about 15% are in second and third place.



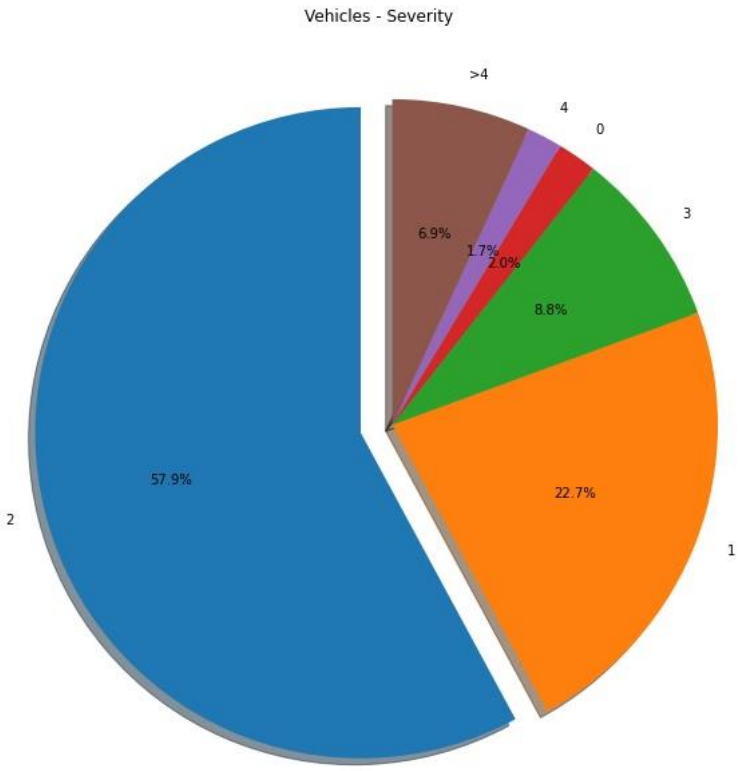
Persons – Severity

About 41% of the accident that resulted in injuries involved two people. Almost 20% of total accidents involve 3 people.



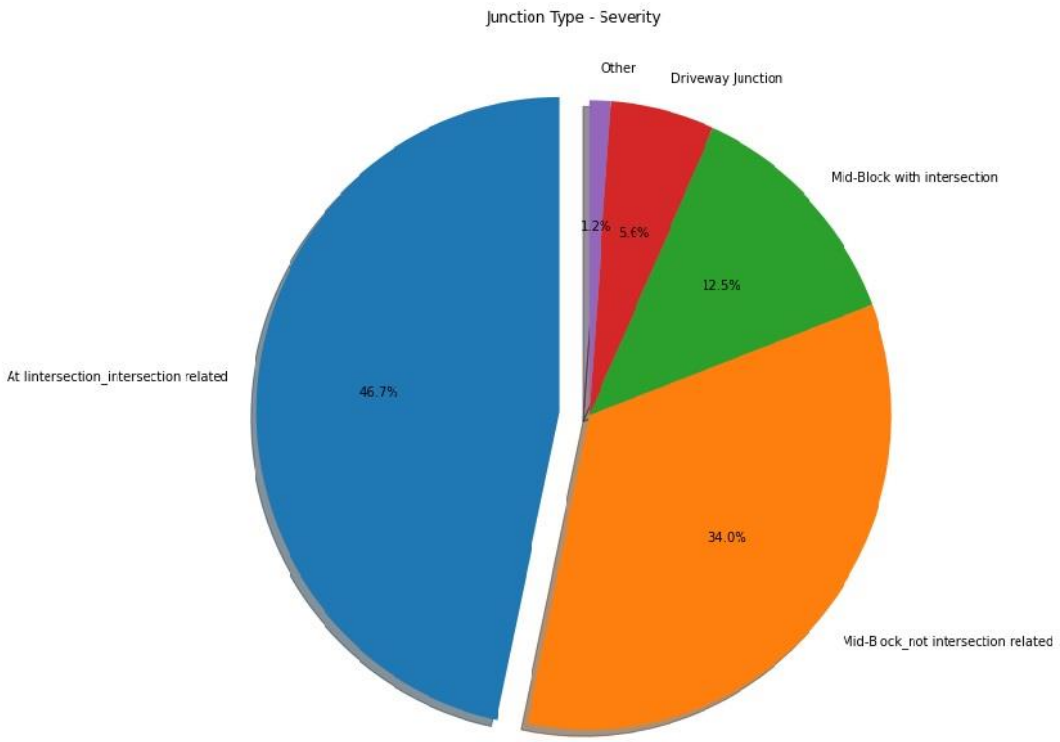
Vehicles – Severity

About 48% of the accidents that resulted in injuries involved two vehicles and a plus of 23% of accidents involve 1 vehicle.



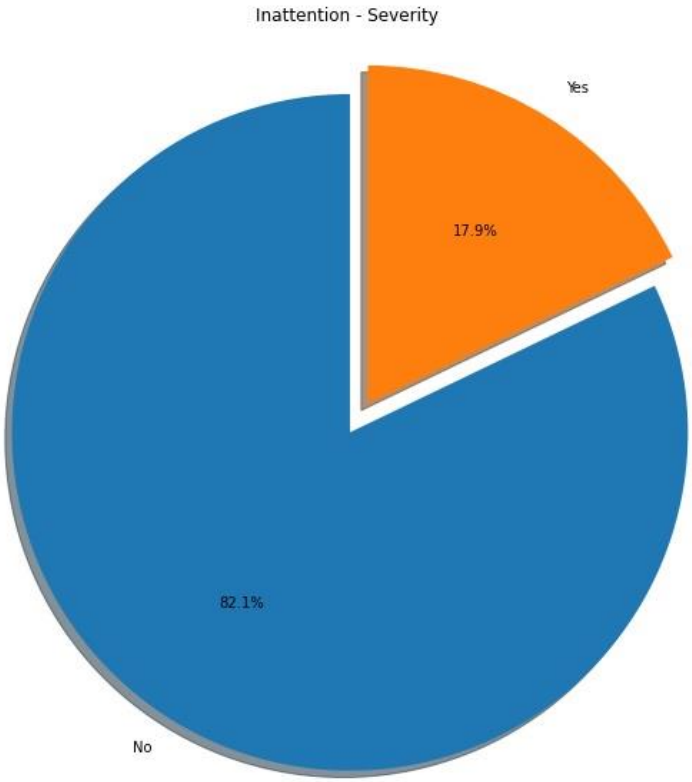
Junction Type – Severity

The most serious accidents have occurred not only at the intersection (47%) but also in the middle block (34%) not related to the intersection. These two cases together contribute to more than 80% of serious accidents.


Junction Type - Severity
Other
Driveway Junction
Mid-Block with intersection
1.2%
5.6%
12.5%
At Intersection_intersection related
46.7%
34.0%
Mid-Block_not intersection related

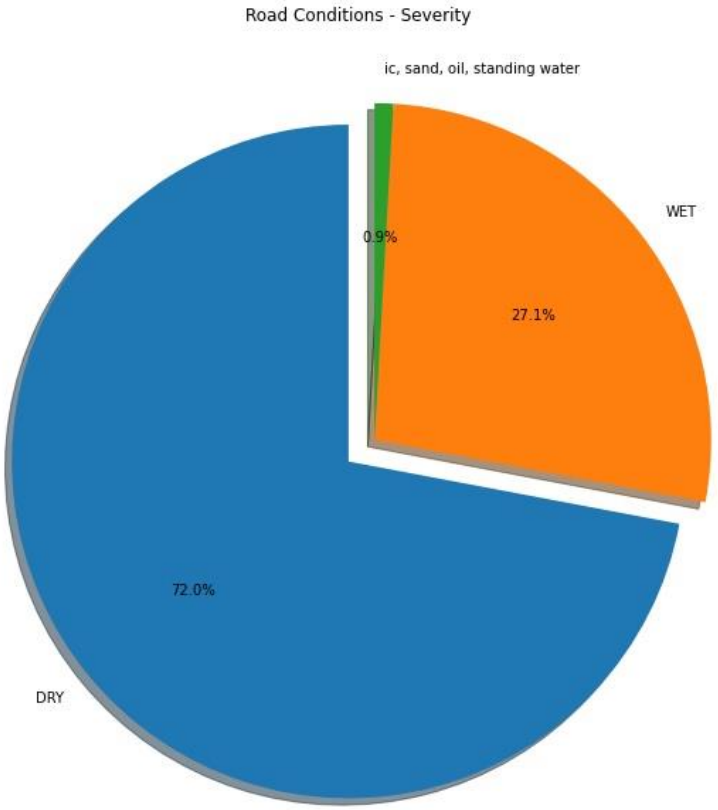
Inattention – Severity

About 82% of serious accidents are not directly related to the carelessness factor, but about 18% of them have occurred immediately as a result of carelessness.



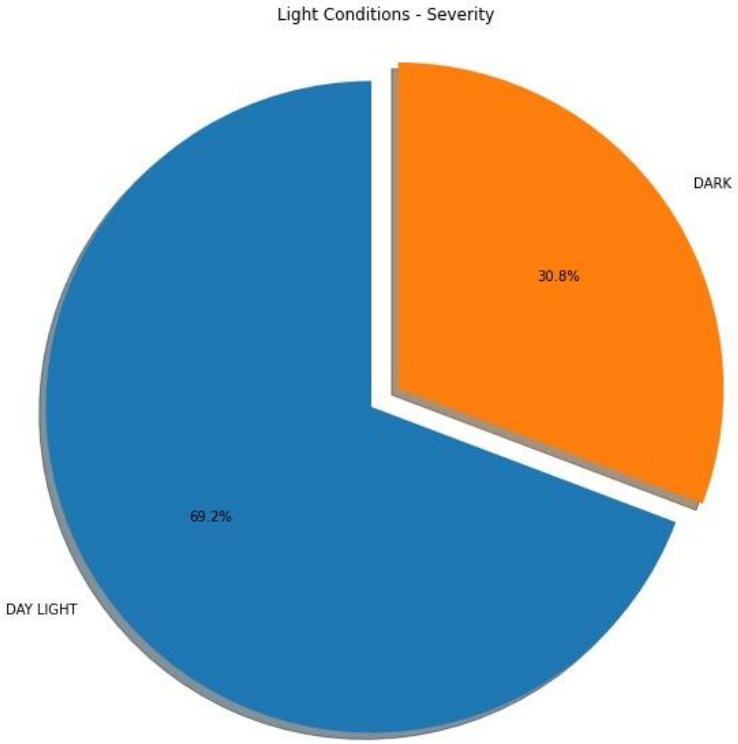
Road Conditions – Severity

DRY conditions are about **72%** while other conditions containing liquid and ice / sand / oil / stagnant water contribute approximately 28%.



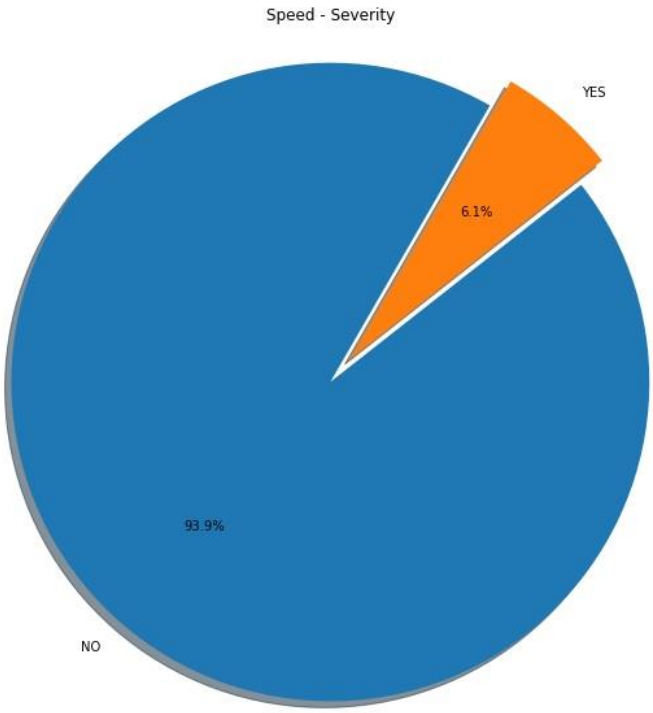
Light Conditions – Severity

About 70% of severely injured accidents have occurred in daylight, about 30% of them have occurred in the dark.



Speed – Severity

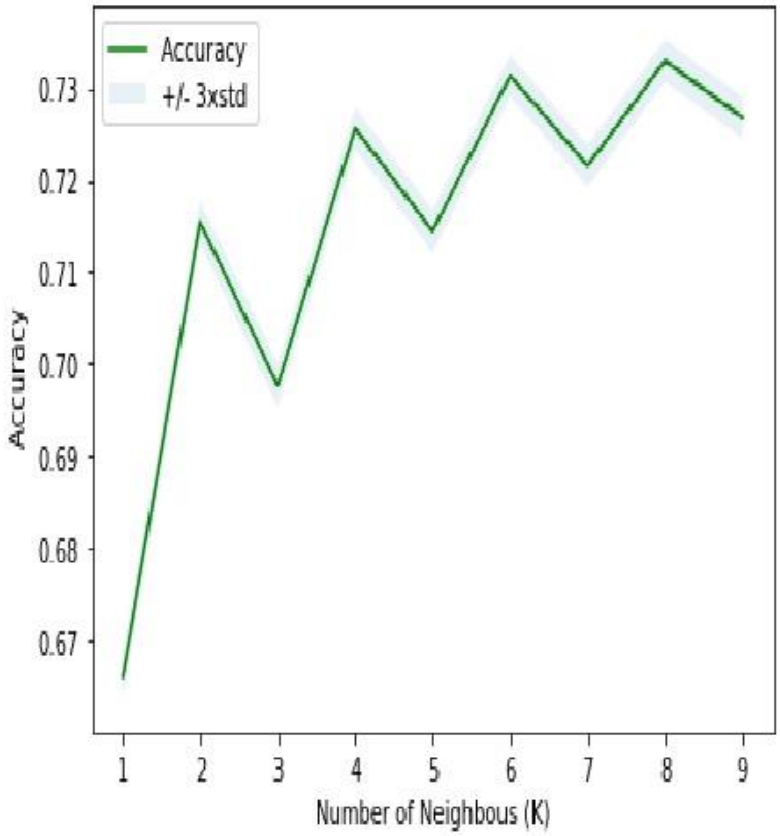
About 6% of In severely injured accidents, acceleration was the direct influencing factor, while in the rest (approximately 94%) factors other than acceleration played major roles.



3.2. Modeling, testing and evaluation

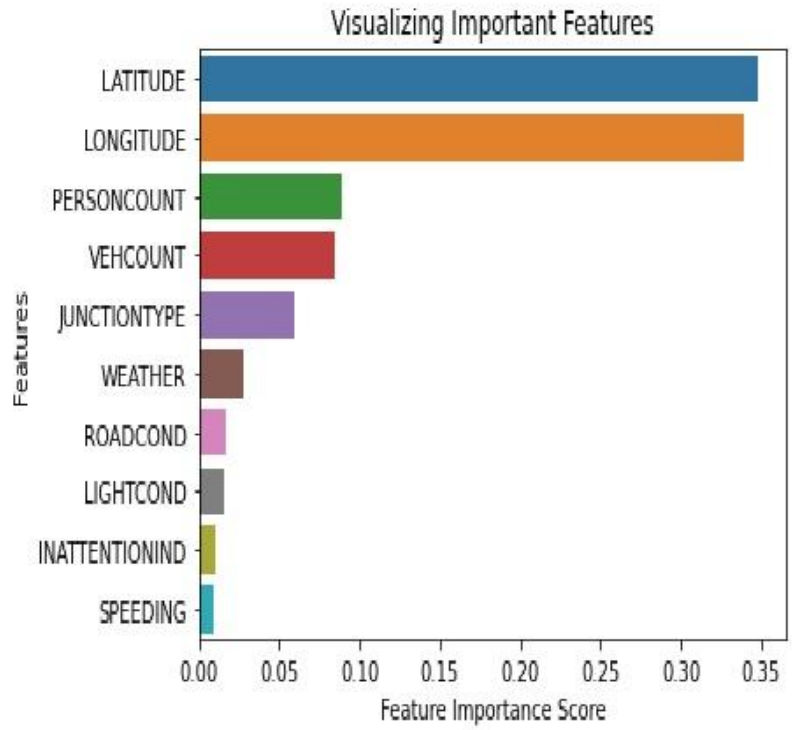
Data are divided into train and test sets, 80% for guidance and 20% for test.

The accuracy of the K-Nearest Neighbors model is examined;



An accuracy test is applied to evaluate the model with a result of 0.73 accuracy for the model. The disadvantage of K-Nearest Neighbors can be mentioned as the point that depends a lot on the quality of the data and since the data set had missing values, this may be the reason why it does not get better accuracy using the K-Nearest Neighbors.

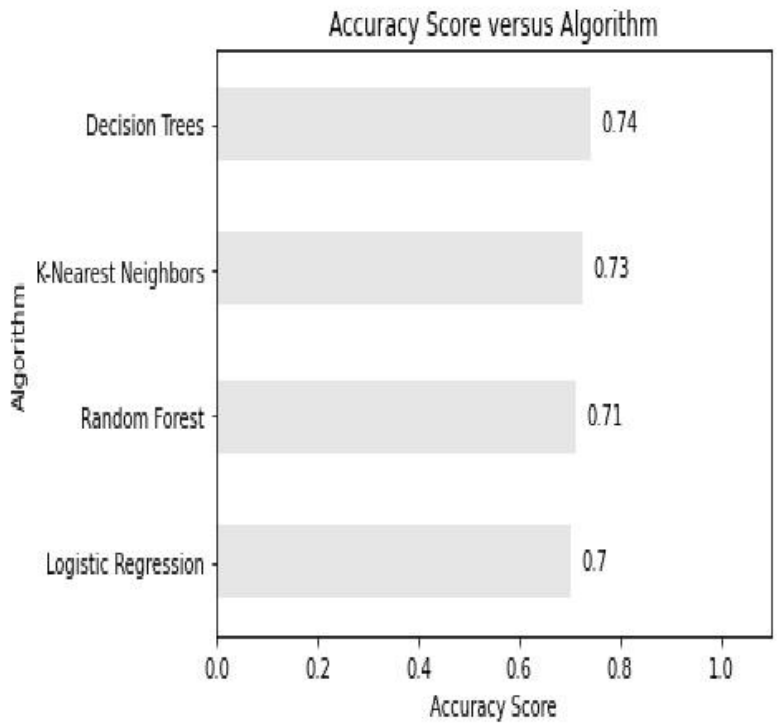
The next algorithm is the Random Forest which resulted in an accuracy of 0.71, slightly lower than the K-Nearest Neighbors and the decision tree. The Random Forest classifier, despite the decision tree, creates many trees to process during training. This can, however, increase the likelihood of over-placement, as noise can also be considered as the actual data points.



4. Results and Discussion

Running the K-Nearest Neighbors classifier took longer than modeling Decision Trees.

The efficiency and compatibility of the decision tree to handle this data set is evident.



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 developing the decision tree machine learning model to improve its accuracy further. Adding more data to the dataset can help to compensate for the missing values. Gathering more data about other parameters such as the age of the drivers can also help to gain a more detailed insight into the car accident severity.

Better models could have been made if data was more comprehensive and had less unknown and missing values. The analyses done could also have held more value if greater target variable class-data was available and was not limited to property damage only and physical injury.

5. Conclusion and Recommendations

The severity of a car accident gives rise to the following causes:

Adverse weather conditions, road conditions and lighting conditions.

Careless driving, violation of the speed limit also contribute to the likelihood of an accident and compete closely with the aforementioned factors. Make sure the driver is relaxed, follows the speed limit and is not under the influence of travel.

It is necessary to develop a holistic management model where:

The number of emergency responders in areas where accidents occur and the reduction of response times to incidents

The existence of health units near the areas where accidents occur

Investments in hotspot accident areas in order to improve road conditions, signage and lighting, traffic management systems.