Car Accident Severity

Capstone Project

Week 2_Assignment

Full Report

1. Introduction / Business Problem

1.1. Background

Road traffic injuries (RTIs) are a major public health problem. The annual global status reports on road safety, launched by the World Health Organization (WHO), highlights that the number of road traffic deaths has exceeded one million in recent years. That is over 3000 people dying on the world's roads every day.

To combat this major problem, the volume of research in the areas of accident analysis and prediction has been increasing over the past few decades. The analytical data mining solutions can be largely employed to determine and predict related influential factors and thus to explain RTIs severity.

Among the analytical data mining solutions, supervised machine learning (ML), has become a popular scientific method to predict the severity of accidents. The reasons for this popularity, can be referred to the capacity present in ML to identify the existing patterns in the data and make predictions via the establishment and evaluation of diverse algorithms. Moreover, the ability of MLs to handle large amounts of data is an additional asset for this purpose, as the data on road traffic accidents are often sparse and largely extended.

1.2. Objective

The objective of this project is to predict the car accident severity by training and evaluating supervised machine learning algorithms with the help of a shared dataset including the record of collisions provided by SDOT Traffic Management Division, Traffic Records Group (timeframe: 2004 to present) for the city of Seattle, a city on the west coast of the United States.

As the approach to reach this objective, the cross-industry standard process for data mining (CRISP-DM) (Figure 1) will be implemented. During the coursework of this procedure, the raw

data will be accordingly well understood and prepared before being fed for the predictive modeling analysis in the next steps.

Figure 1, Cross-industry Standard Process for Data Mining (CRISP-DM)[https://en.wikipedia.org]

1.3. Interest

The report of this project can be targeted to stakeholders, who are involved with road traffic injuries incl. road administrators, traffic control authorities, and emergency road services in order to help them predict the car accident severities and improve the road users' safety margins.

2. Data

2.1. Data Source

Followed by the Cross-industry Standard Process for Data Mining (CRISP-DM), it is now time to understand the data and then prepare it to be fed into the modeling tools. The given dataset used in this project (provided by the SDOT Traffic Management Division, Traffic Records Group, timeframe: 2004 to present, Seattle, United States) can be downloaded here.

The shared dataset (Data-Collisions.csv) contains 194673 and 38 columns including the labeled data. The labeled data is the "Severity Code", which describes the fatality of an accident. In the shared dataset, the severity code column consists of two values: 1 for property damage and 2 for injury. The dataset includes different attributes, describing a variety of conditions e.g.

location, weather, light, road, collision types, and so forth that may influence the severity of the accidents. The attributes are of the types of int64, float64, or object.

2.2. Data Balancing

In the given dataset, the targeted variable (Severity Code) has more observations in one specific class (1, corresponding to approx. 70% of the collision severity accompanied by property damage) than the other (2, corresponding to approx. 30% of the collision severity accompanied by injury). Figure 2.a shows the initial imbalanced dataset in this work.

This imbalanced dataset needs to become firstly balanced, otherwise, the models that will be developed later will be biased. Resampling is a widely adopted technique to address this issue. It consists of randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm (under sampling) or randomly duplicating observations from the minority class to reinforce its signal (over sampling).

By under sampling, a large amount of data, which can be later used for the prediction of severity will be lost. Accordingly, the oversampling technique is preferred and implemented in this work. Figure 2.b shows the resulting balanced dataset, combining the majority class with over sampled minority class.

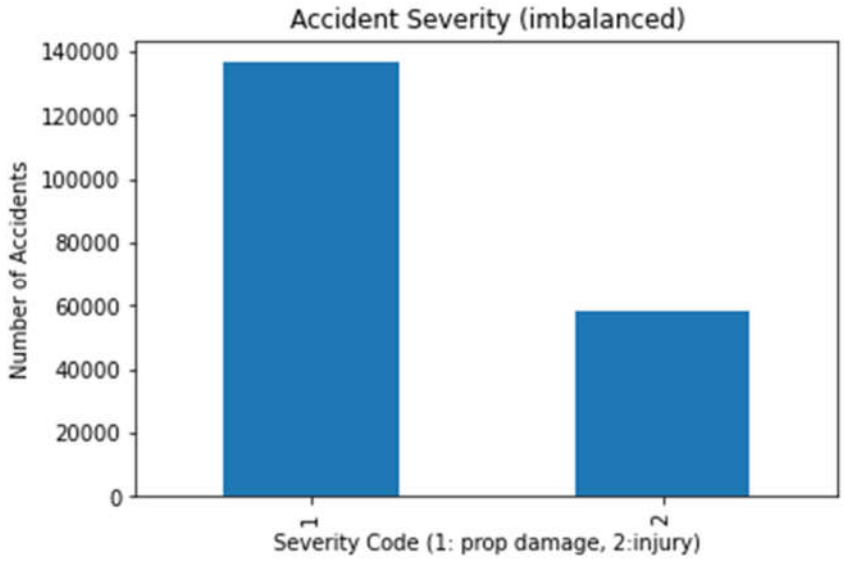

Figure 2.a, Imbalanced Dataset

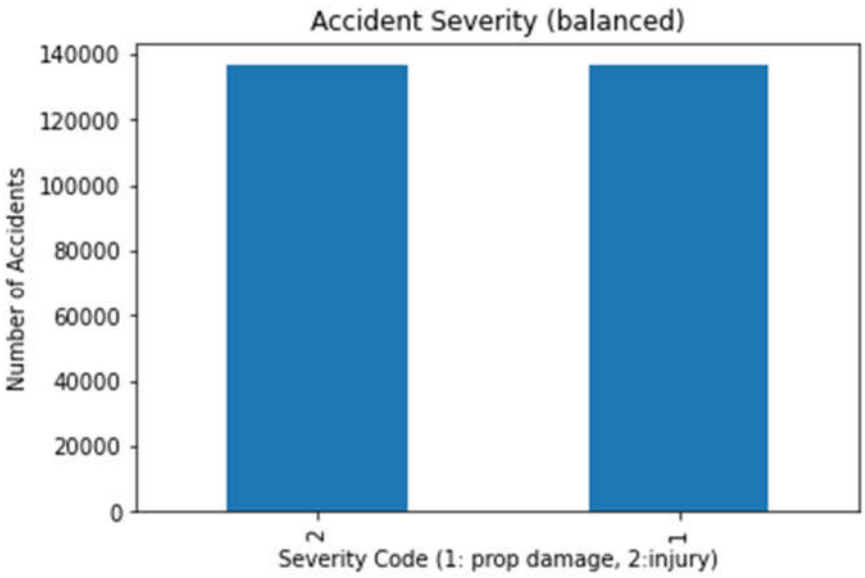

Figure 2.b, Balanced Dataset

2.3. Data Cleaning

There are several issues which are needed to be addressed during the data cleaning. One issue is many cells with missing values. The other issue with these missing values is that they are widely spread within 19 columns out of 38 columns in the dataset coming with a “NaN” mark. As this distribution ratio is considerably high, the replacement of the missing data with reasonable new values is a better option as far as possible.

The other issue is the presence of both numerical and categorical data in the dataset. To this effect, the replacement is done by the frequency for the categorical variables and by mean for the numerical values. The missing categorical values that are replaced by the largest frequency belongs to the columns WEATHER, SPEEDING, LIGHTCOND, ROADCOND, JUNCTIONTYPE, INATTENTIONIND, COLLISIONTYPE, and ADDRTYPE. The missing numeric values in columns X and Y are replaced by the mean of the belonging columns, respectively.

What should be also considered, specifically for processing the data in the next steps, is the incompatibility of categorical variables with the predictive model analysis tools. For example, to develop regression models and being able to use packages such as Sklearn, these variables are converted into indicator variables during the data cleaning after handling the missing data.

2.4. Feature Selection

Taking a closer look into the dataset reveals that many of the columns contain inter-organizational codes which are not relevant to the case of this study and are deleted. These columns include OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY,

EXCEPTRSNCODE, SDOT_COLCODE, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, and CROSSWALKKEY. For example, the column SDOT_COLCODE refers to the codes given to the collision by SDOT or the columns INCKEY and COLDETKEY contain the ESRI unique identifier and so on.

Some of the columns also consist of redundant or not enough useful information. For example, there is a second SEVERITYCODE.1 in addition to the SEVERITYCODE which will be deleted. The redundancy in data is also observed for columns such as EXCEPTRSNDESC with no description. The other example is the column UNDERINF which addresses the question:" Whether or not a driver involved was under the influence of drugs or alcohol?" There is however another column named INATTENTIONIND addressing the question: "Whether collision was due to inattention?"

The level of attention in people is usually decreased upon consuming drugs or alcoholic drinks. Accordingly, UNDERINF is deleted and INATTENTIONIND is remained to count for the level of attention. The same analogy is considered for the column LOCATION, as both the X (longitude) and Y (latitudes) are given. Working with X and Y coordinates has also the advantage of a more precise description of the places where the accident occurred. For more clarity, X and Y are also renamed to LONGITUDE and LATITUDE. The same analogy is also considered for the columns STATUS, INCDATE, INCDTTM, SDOT_COLDESC, PEDROWNOTGRNT, ST_COLDESC, PEDCYLCOUNT, HITPARKEDCAR, SEVERITYDESC, ADDRTYPE, COLLISIONTYPE, and PEDCOUNT. The 10 features selected at the end of this step are listed in Table 1.

Table 1, **List of features being selected in the feature selection**

	Feature	**Description**
1	LONGITUDE	longitude
2	LATITUDE	latitude
3	PERSONCOUNT	total number of people involved in the collision
4	VEHCOUNT	the number of vehicles involved in the collision
5	JUNCTIONTYPE	category of junction at which collision took place
6	INATTENTIONIND	whether or not collision was due to inattention
7	WEATHER	a description of the weather conditions during the time of the collision
8	ROADCOND	the condition of the road during the collision
9	LIGHTCOND	the light conditions during the collision.
10	SPEEDING	whether or not speeding was a factor in the collision

3. Methodology

In this section, an overview of the methodology employed in this work is discussed. After the features are selected, they are applied for an explanatory data analysis to figure out more about their effects on the severity of the accidents. The focus is on identifying the feature conditions that have a bigger effect on the severity which leads to injuries. To do so, the data set is filtered and the values of features that correspond to this level of severity are sorted.

The latitude and longitude are firstly paired as the coordinates of accident locations and are assigned to a new data frame. The Folium package is imported, and looping is done in which a feature group plus a circle marker is employed to gain a map of Seattle with the points representing the locations of the accidents.

After the visualization, data exploration is followed by investigating the proportions of other features. Pie charts are used to display the numeric proportions by dividing a circle into proportional slices. The different conditions of weather (clear, raining, overcast, and other) are firstly investigated to figure out which one and to how much percent has mostly contributed to severe accidents. The number of persons (o, 1, 2, 3, 4, 5, 6, and more) involved with the accident is then studied. Subsequently, the effect of vehicle count (0, 1, 2, 3, 4, and more) is studied.

The data in the dataset also shows the effect of junction types on the accident severity. The accidents have occurred either directly at intersections or in mid-blocks and drive-way junctions. A pie chart is drawn to determine the importance of each type. Afterward, the effect of inattention is analyzed. The next pie deals with the question that, how extreme the effect of daylight or darkness (due to dusk, dawn, and off streetlight) is. The effect of road conditions comes next in which different conditions including wet, ice, sand, oil, and standing water will be studied. To address the question that, how the effect of speed on the accident severity is, a pie chart analysis is also drawn. All the figures of this explanatory data analysis are gathered in the **Exploratory Data Analysis Section**.

In the next step, the features are processed for predictive modeling analysis, in which various supervised machine learning algorithms are checked to build the model. For this purpose, the data is normalized, and after the train-test-split, classification techniques are implemented. The first algorithm being implemented is the K-Nearest Neighbors (KNN). The KNN works based on the classification concerning the K nearest points in the vicinity of the point to be predicted. Since the KNN algorithm is robust regarding the search space and therefore the classes do not have to be linearly separable, it is chosen in this work.

The next algorithm is the decision tree, in which every internal node is a representation of a test on an attribute, each branch is the representation of the outcome of the test, and each leaf node is the representation of a class label. This flowchart-like technique is famous for its high accuracy and is less sensitive to missing values and non-scaled data. Accordingly, it is used for the classification in this work.

The other algorithm used in this work is logistic regression. It passes the input through the logistic/sigmoid but then treats the result as a probability. This algorithm has good training abilities and its relatively good capacity for avoiding overfitting issues are the reasons for being employed in this project. The last selected classification technique is Random Forest. Random Forest works based on constructing a multitude of decision trees and because of its capacity to judge the interaction between different features (to come out with higher dimensional features data) it is chosen.

Random Forest technique also gives the possibility to get an overview of the most important features. To get the benefit of this possibility, after being done with the modeling using the Random Forest classifier, this classifier is employed for giving an estimation of the most important features.

Once the training and testing of the above-mentioned algorithms are done, they are tested and evaluated using the accuracy measure, by the calculation of the number of correct predictions and the total number of predictions ratio. The results of the predictive modeling analysis and related tests are gathered in the **Modeling, Testing and Evaluation Section**.

4. Exploratory Data Analysis

The marked map is depicted in Figure 3 for a limit of 300 data points. It is programmed so that it is possible to increase/ decrease the number of markers on the map in case of necessity.

The results of the pie chart for the weather are depicted in Figure 4. As the results show, around 65% of severe accidents with the injury have occurred in the clear conditions in Seattle. Raining with around 19% and Overcast with around 15% are in the second and third places, respectively.

The effect of person count is depicted in Figure 5. As the figure shows, in around 41% of the accident leading to injuries, there were two persons involved. In nearly 20% of total accidents, 3 persons were involved. The pie chart of the person count is followed by the pie chart of the vehicle count (Figure 6).

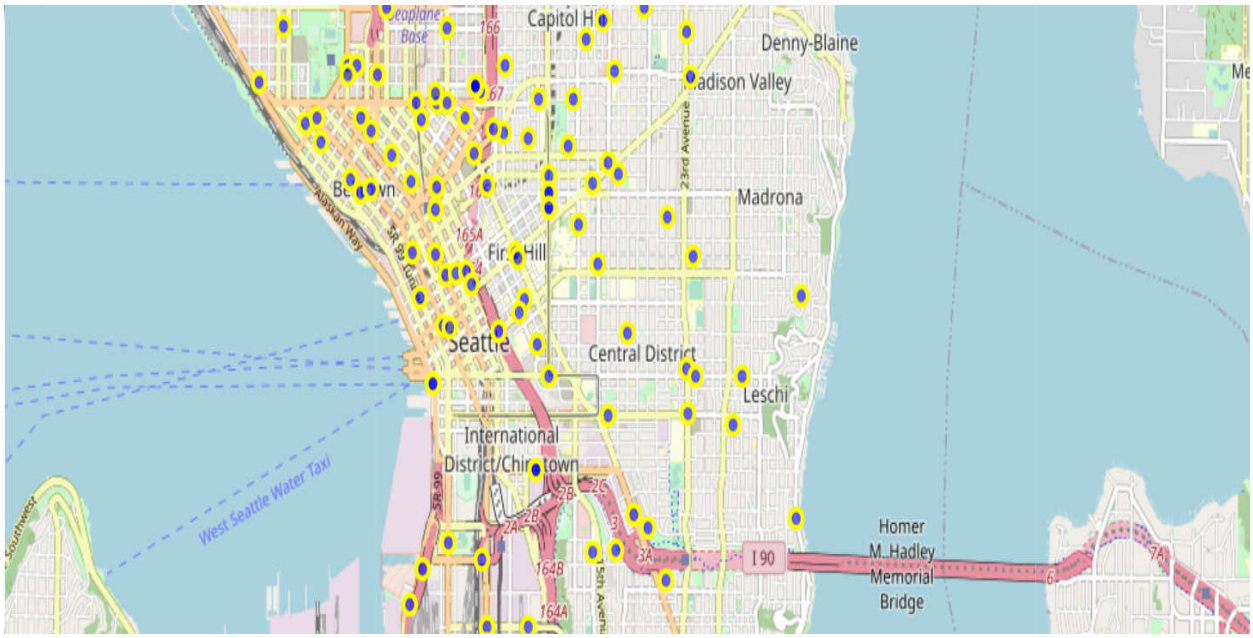

Figure 3, The map with markers of the accident locations in Seattle

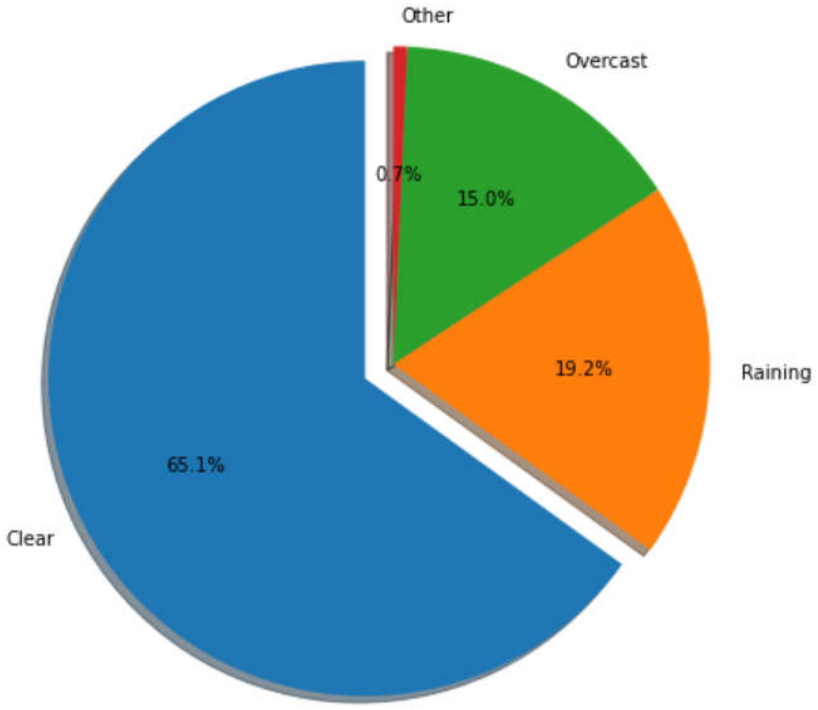

Figure 4, Relationship between the Weather Conditions and the Accident Severity with Injury

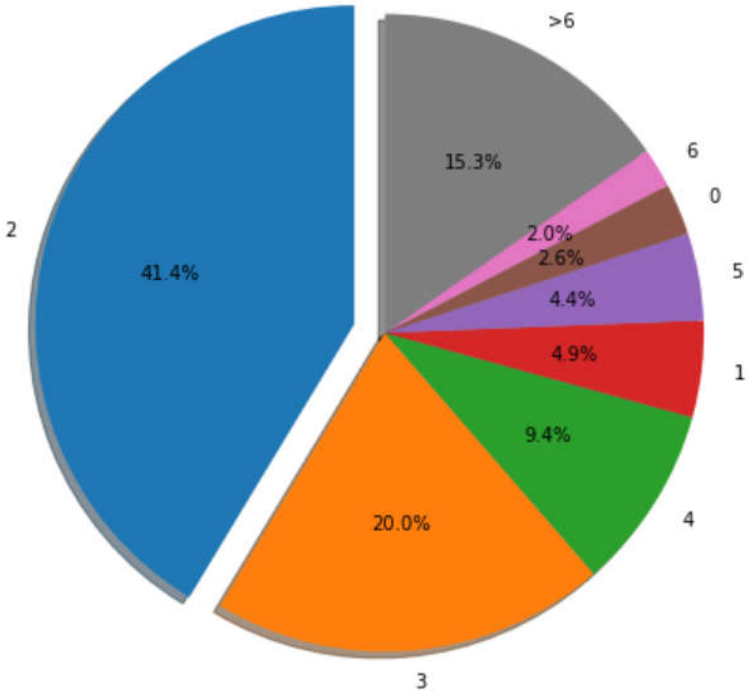

Figure 5, Relationship between the Person Count and the Accident Severity with Injury

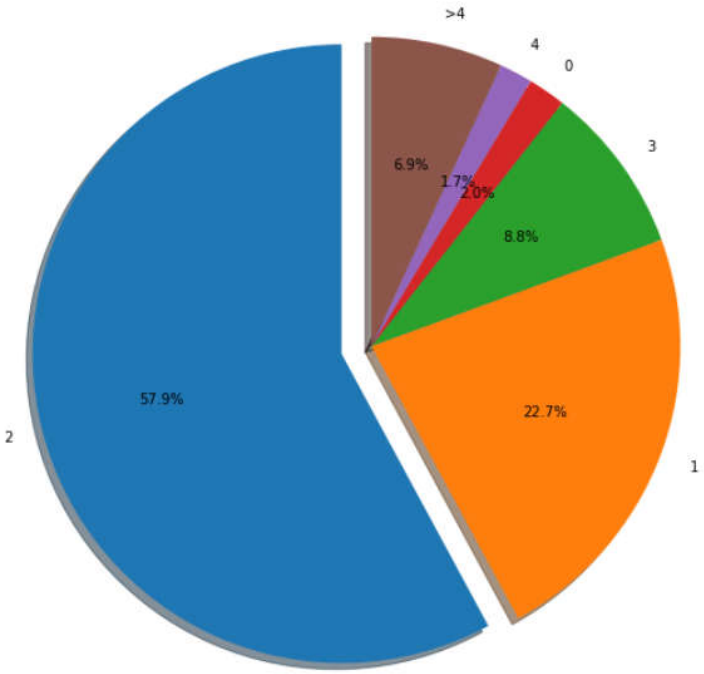

Figure 6, Relationship between the Vehicle Count and the Accident Severity with Injury

The pie chart in Figure 7, shows that most severe accidents have not only occurred at the intersection (which are clearly intersection related) but also at mid-block which are not

intersection related. These two cases together contribute to more than 80% of accident severities.

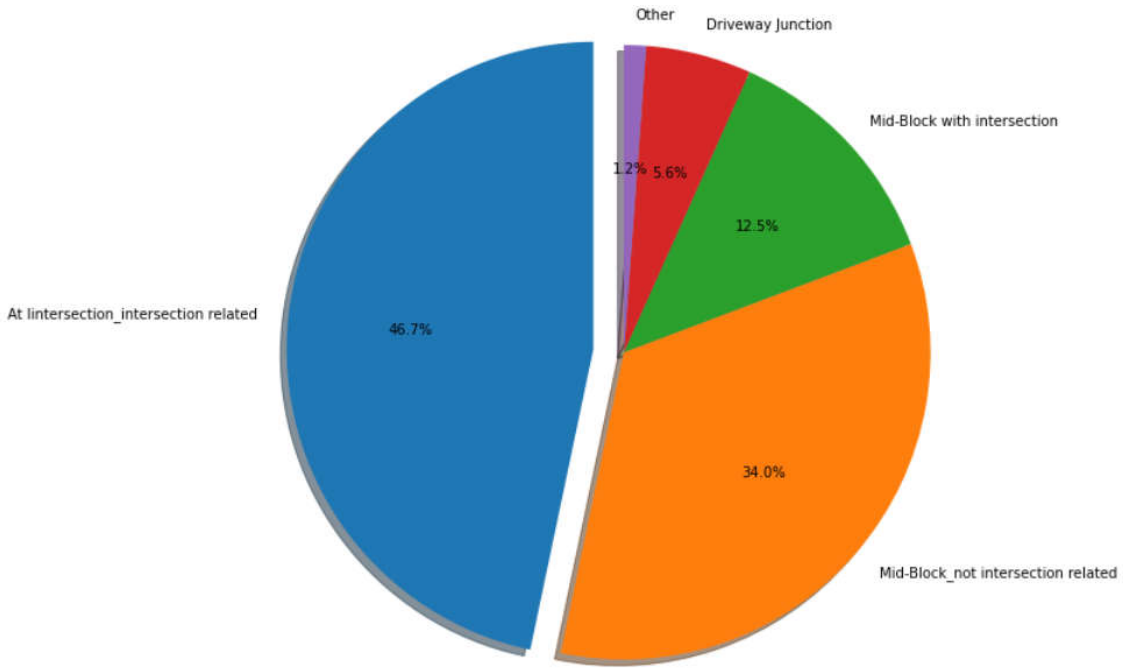

Figure 7, Relationship between the Junction Type and the Accident Severity with Injury

The pie chart of inattention (Figure 8) shows that around 82% of the severe accidents are not directly related to the inattention factor, but around 18% of them have occurred directly as the result of inattention.

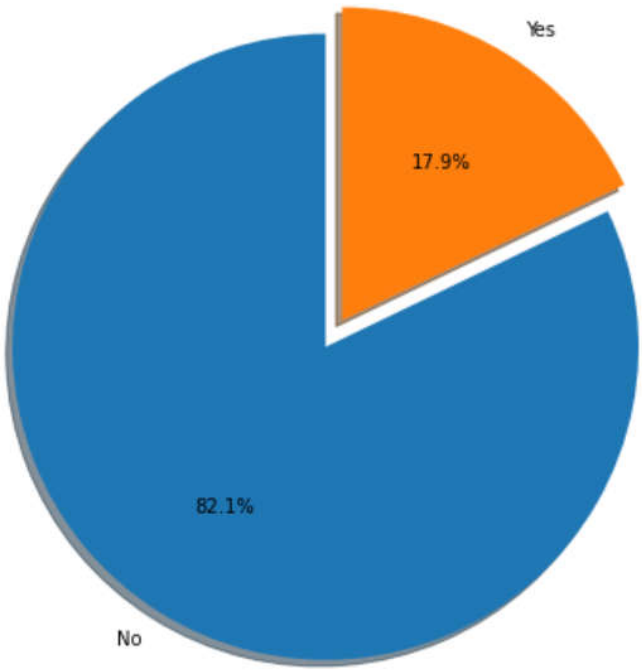

Figure 8, Relationship between the Inattention and the Accident Severity with Injury

The pie chart in Figure 9 shows the relationship between the road condition and the accident severity. The proportion of the dry conditions is around 72% while other conditions containing wet and ice/sand/oil/standing water contribute to approx. 27% and 1% respectively. While around 70% of severely injured accidents have occurred in the daylight, around 30% of them have occurred at darkness (dawn, dusk, night, etc.). (Figure 10)

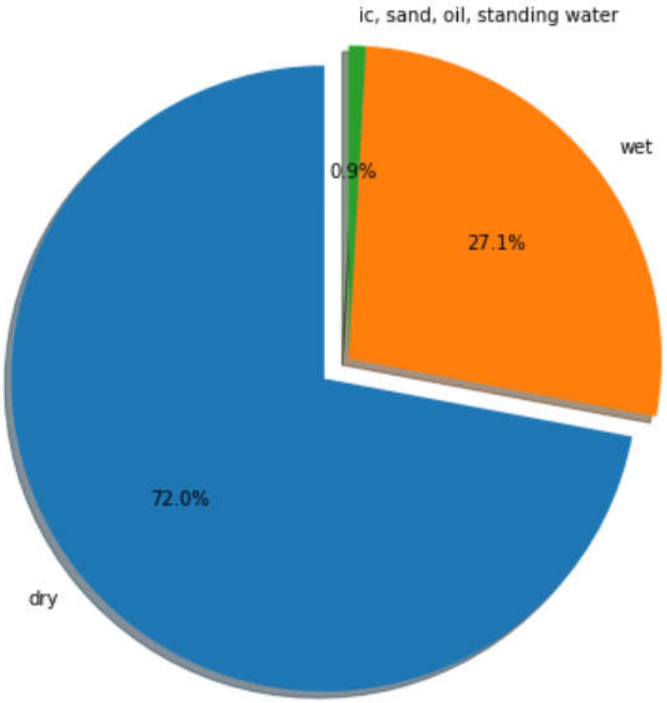

Figure 9, Relationship between the Road Conditions and the Accident Severity with Injury

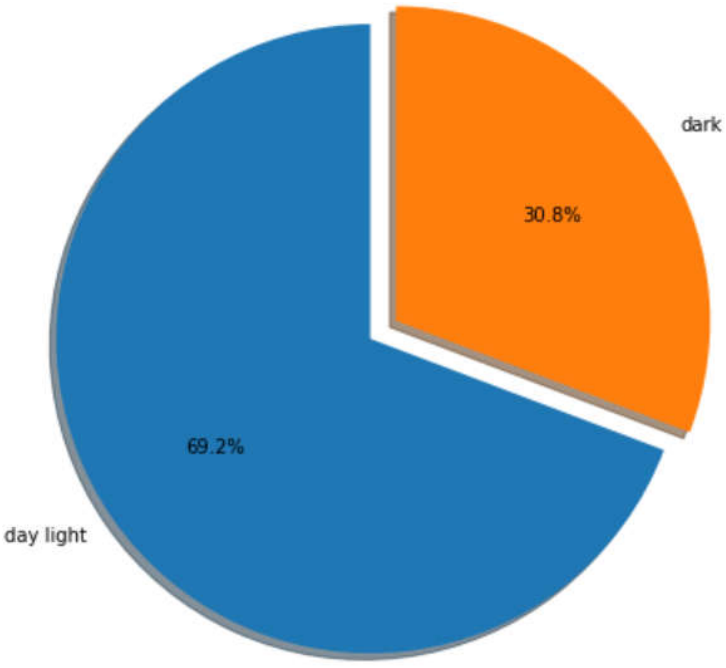

Figure 10, Relationship between the Lightness and the Accident Severity with Injury

The pie chart in Figure 11, shows that for around 6% of the severely injured accidents the speeding was the direct influencing factor, while in the rest (around 94%) other factors rather than speeding have played big roles.

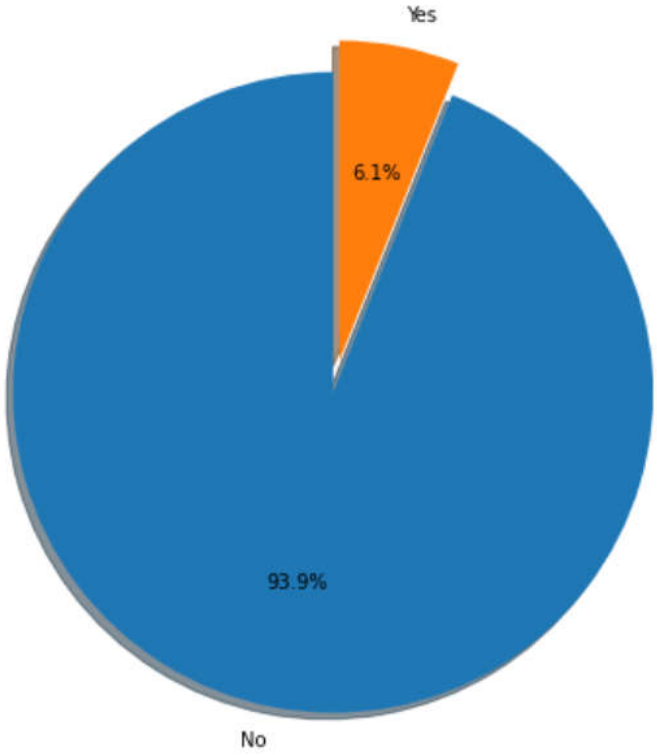

Figure 11, Relationship between the Speeding and the Accident Severity with Injury

5. Modeling, Testing and Evaluation

For the implementation of the machine learning algorithms, the data is divided to train and test sets, 80% for the train, and 20% for the test. To find the best possible option for K, the accuracy of the K-Nearest Neighbors model is examined, resulting in Figure 12. As it is observable from Figure 12, the optimized value of K is obtained as 8.

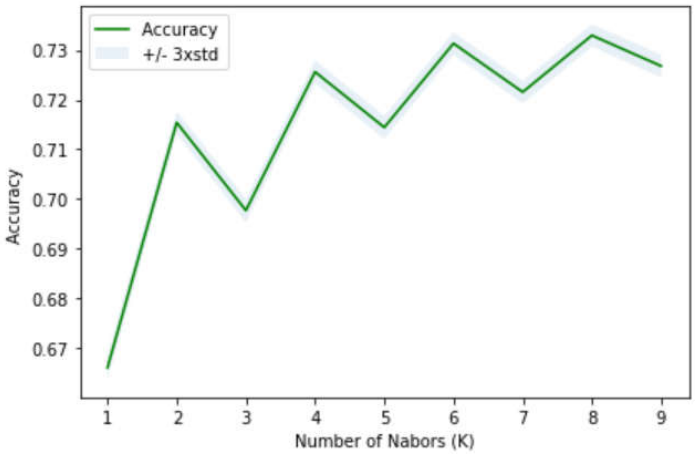

Figure 12, The Model Accuracy for Different Numbers of Neighbors

The model is then rebuilt considering this value and the prediction is done afterward. To evaluate the model an accuracy test is applied resulting in 0.73 accuracy for the model. The disadvantage of K-Nearest Neighbors can be referred to as the point that it is very dependent on the quality of the data and as the dataset had missing values, this may be the reason for not getting better accuracy using K-Nearest Neighbors.

In the next step, a model for the decision tree is developed. The evaluation obtained for the decision tree model is 0.74, a bit better than the accuracy of K-Nearest Neighbors. This better accuracy can be also interpreted as better compatibility of the decision tree to face with a non-initially well-scaled dataset.

Using logistic regression, a lower accuracy of 0.7 is achieved. The reason for getting worse accuracy applying logistic regression can be because of its simplified assumptions (linearly separable) that it makes between the dependent variable and the independent variables.

The next algorithm is Random Forest which resulted in an accuracy of 0.71, a bit lower than K-Nearest Neighbors and the decision tree. The Random Forest classifier, despite the decision tree, creates many trees to process them during the training. This can however increase the probability of overfitting, as noises may be also considered as the actual data points. Applying the Random Forest classifier, the importance of features is visualized in Figure 13.

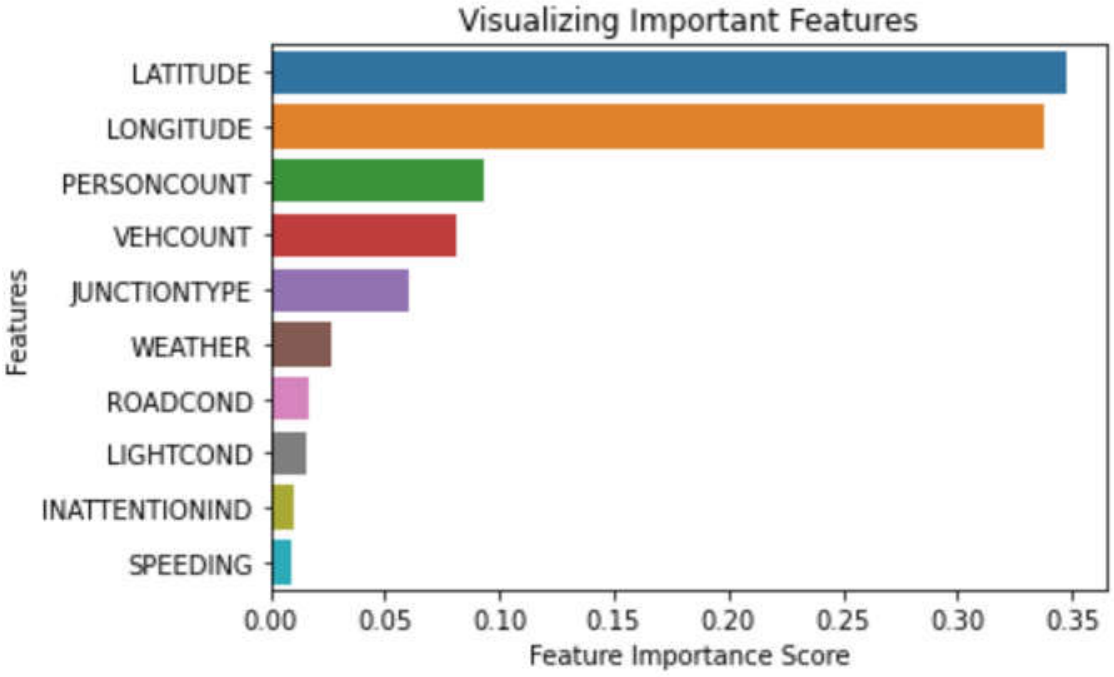

Figure 13, Feature Importance Score based on Random Forest Classifier

6. Results and Discussion

During the modeling with the 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 handling this given dataset.

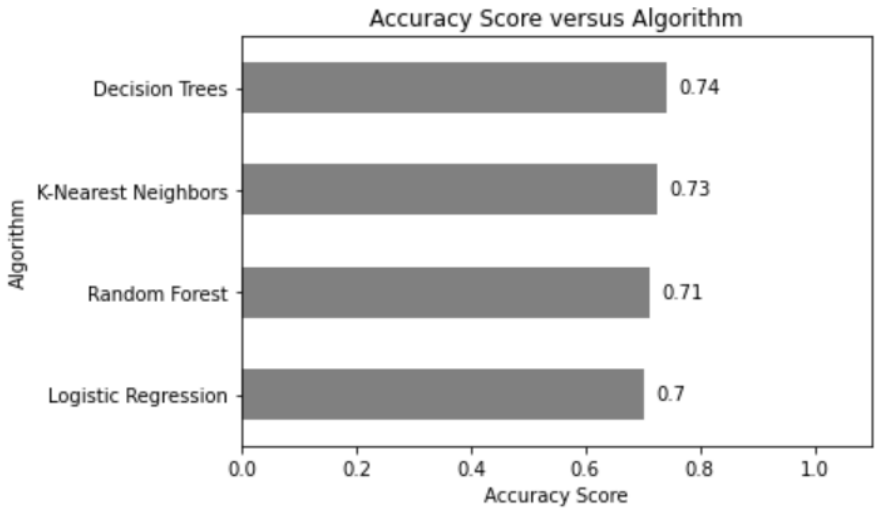

Figure 14, Accuracy Score vs Algorithms

7. Conclusion and Outlook

In this study, supervised machine learning is employed to predict car accident severity. The imbalanced dataset is firstly balanced, and the raw data is understood and prepared in different steps to be used for the predictive modeling analysis. In parallel, an explanatory data analysis is done to gain more insight 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 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.