CAPSTONE PROJECT SEATTLE CAR COLLISIONS

Adrián Arranz Mezquita

Index

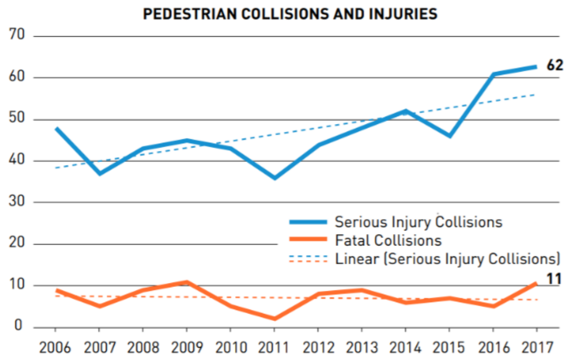
1. Business Understanding

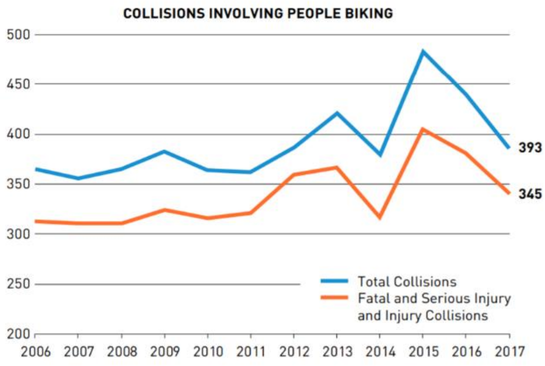
1.1. Problem

Car accidents are a worry for the whole world. As city population increases day by day, so does the number of cars in cities and so does the pollution. That is why the last years one of the politics campaigning spots had been new measures to reduce $C0_2$ emissions. Cities like Seattle are having success in promoting greener solutions of transport (promoting bike use, investing on public transport infrastructure and promoting pedestrian areas, among others), but now they need to assure that are safe. The aim of these project is to study the possible causes related with pedestrian-involved and bike-involved accidents, identifying the measures that need to be taken to reduce collisions amount and its severity and, finally, to develop a predictive model to evaluate severity.

1.2. Background

The study will be focused on Seattle traffic data. According Seattle Department of Transportation, while city population is growing by a 1.8% as yearly rate from 2011, average daily traffic remains stable. Which means a new trend in transportation usage, clearly reducing the use of cars, while the rates of transit ridership and going-by-walk have not stopped of growing (Transit ridership presents a 14.3% yearly growth average, followed by Pedestrian with an 8.3% yearly growth average from 2011). However, bike usage has not register remarkable growth during past years, neither facilities have been yet deployed for them. This drive us to an increase of pedestrian collisions by 9.5% by year. Also bike-involved collision has raised from 2011 to 2017 by 1.5% yearly growth average, even if not related with any traffic increase. This insights make us aware that for having success implementing future measures to reduce pollution in cities, first we need to think in new parallel measures for increase safety for these new users.





1.3. Stakeholders

Reduction in severity and amount of collisions is interesting for Public Development Authority as for the Department of Transport and Departments of Environment, to foster their policies.

2. Data Understanding

2.1. Data Description

The dataset provided by the course has a total of 38 columns x 194,673rows. There is a column called Severity Code, with specific values 1 or 2, meaning property damage-only collision or injury collision. By developing a machine learning model we will try to predict these value, based on other features given as location, day, weather, road condition or light condition.

As the dataset show duplicated columns, blank cells and irrelevant data, first step is data cleaning.

2.2. Data Cleaning

There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there. These columns included pedestrian granted way or not, segment lane key, cross walk key and hit parked car.

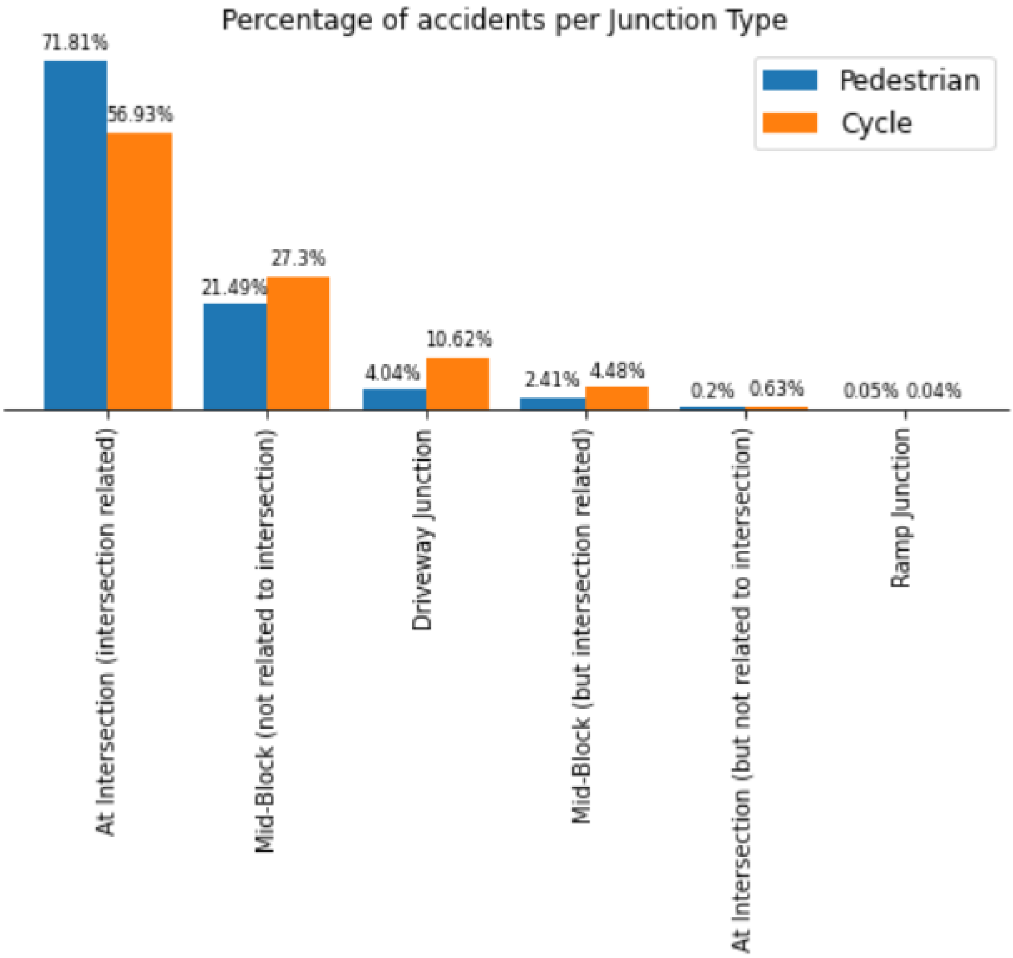
For the first part of the project, as I will analyze just the cycle and pedestrian-involved accidents I will use a different data frame that for the machine learning model training. If I were using just the data frame studied in the exploratory analysis the predictive model would not be good at all because of the heavy unbalance of the severity data. The features that we will use for the exploratory analysis will be:

Feature Variable	Description
X	Latitude Coordinates
Y	Longitude Coordinates
INATTENTIONIND	Whether the driver was or not inattentive (Y/N)
UNDERINFL	Whether or not the driver was under alcohol influence (Y/N)
WEATHER	Weather condition during collision
ROADCOND	Road condition during the collision
LIGHTCOND	Light condition during the collision
SPEEDING	Whether the car was above the speed limit at the time of collision
JUNCTIONTYPE	Whether the collision has been taken into an intersection, ramp…

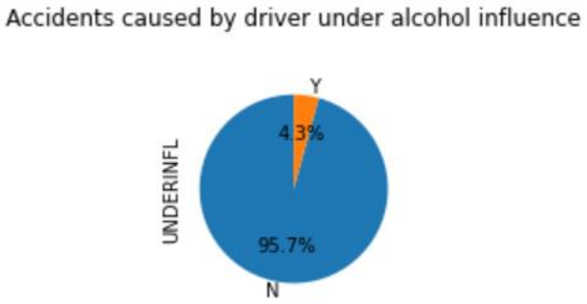
When we analyze de influence of each variable, first we need to check if there are NaN values, to remove them, or if there are any inconsistencies within the data, for example data formatting or data normalization.

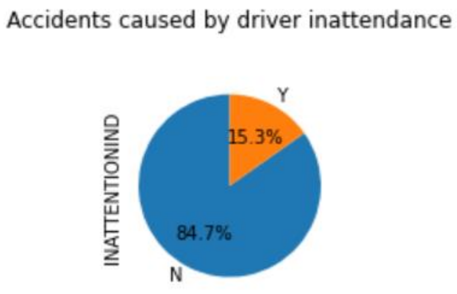
3. Exploratory Data Analysis

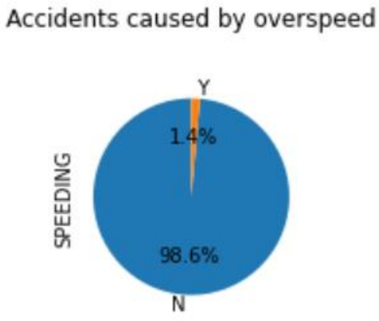
During this part of the work, I am going to analyze the data aiming to find or to disprove some general assumptions. Hence, I start with the analysis of pedestrian and bicycle-involved accidents regards to junction types in where they have been taken place:



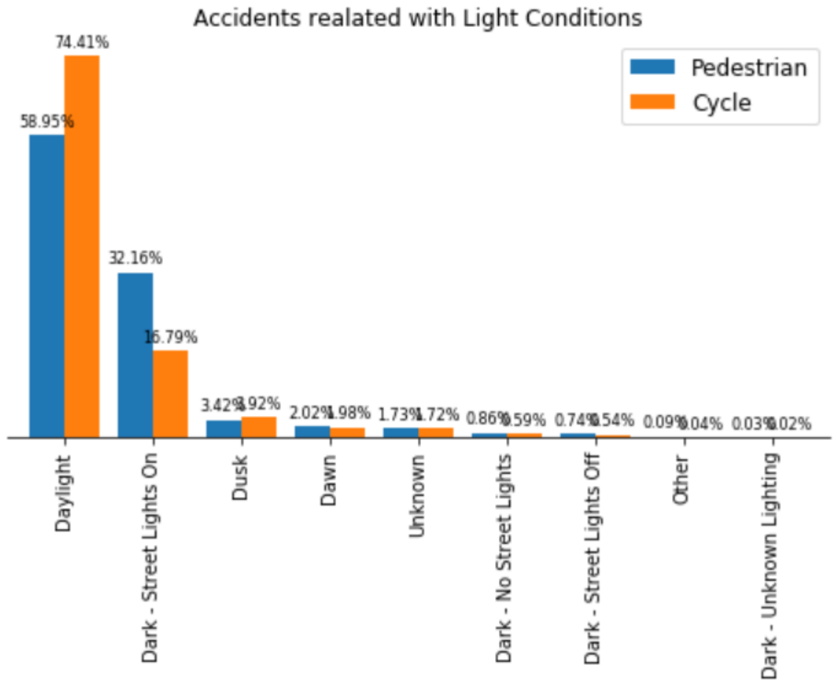
For both, pedestrian and cycles, the majority of accidents take place in intersections, especially for pedestrian raising up to almost 72% of accidents, while cycle-involved accidents are more spread. It is understandable as bikes need to use sometimes car driveways when no specific way exists for them. The data that worries me most is the high percentage of mid-block not related to intersection, which spots not proper behavior by pedestrian or cycles crossing away where it is not allowed. Later, we will check the zones where more accidents take place, so one of the options to take by the government could be to check if there are enough pedestrian crosses within those areas. To analyze root causes from intersection related accidents we will need to acquire more related data as if there are a traffic lights, their status, etc. We could also think that they could be related with driver status: inattentive, under influence or overspeed. But according with the graphs below the percentage of drivers fulfilling any of this conditions is not high enough to conclude it as the problem. However, inattentive drivers is about the 15%, so working on reducing this number could help shorting severity an overall amount of accidents.

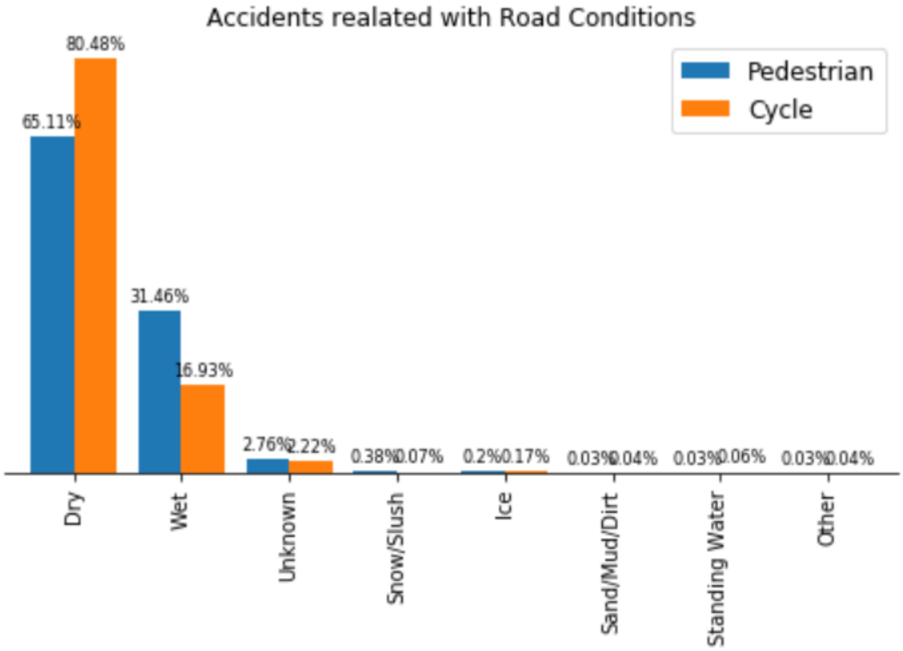


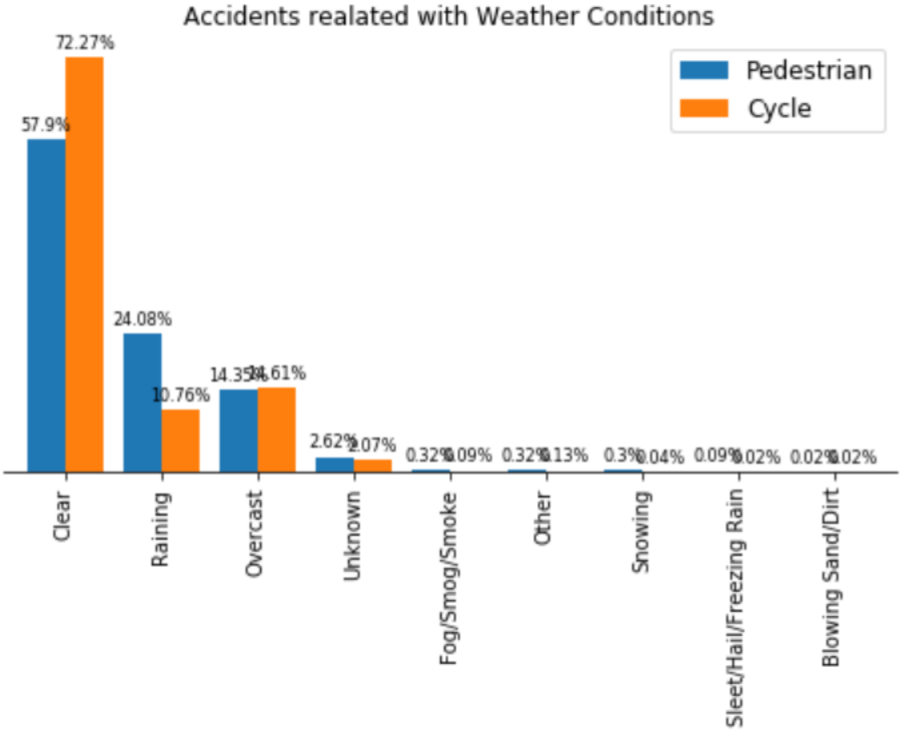




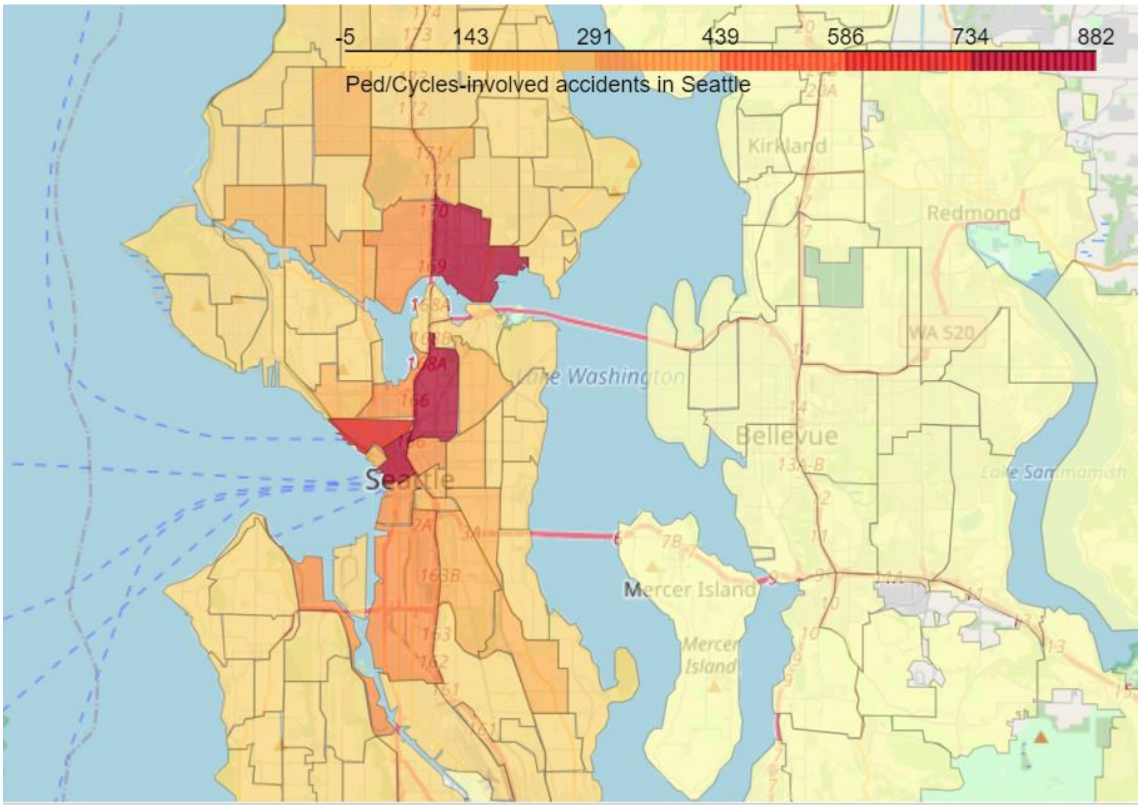
Next thing I think to analyze is whether or not accidents can be related with weather, road conditions or light conditions. By looking at the graphics below, we can see at a glance that most of accidents take place with clear weather, with dry conditions of the road and during daylight. However, for pedestrian-involved accidents percentages do not raise level so high, so we can confirm that night accidents, wet road condition accidents and raining/overcast accidents have an important height, over 30%. Taking into account that in Seattle there are 156 days of rain in a year, we need to further investigate what is happening these days, because maybe there are flood in some areas or there are general problems in visibility to be solved.



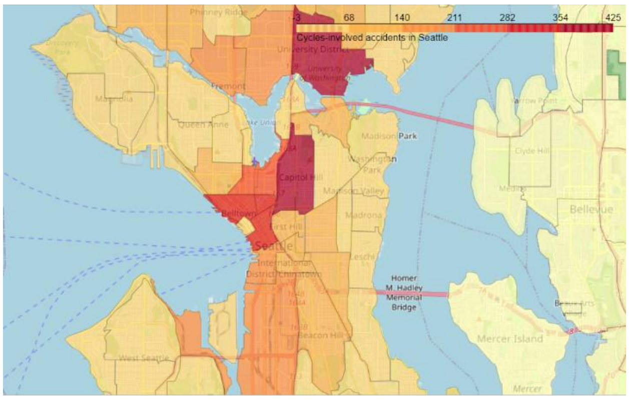

Accidents realated with Road Conditions
Pedestrian
Cycle
65.11%
80.48%
31.46%
16.93%
2.76%
2.22%
0.38%
0.07%
0.2%
0.17%
0.03%
0.04%
0.03%
0.06%
0.03%
0.04%
Dry
Wet
Unknown
Snow/Slush
Ice
Sand/Mud/Dirt
Standing Water
Other

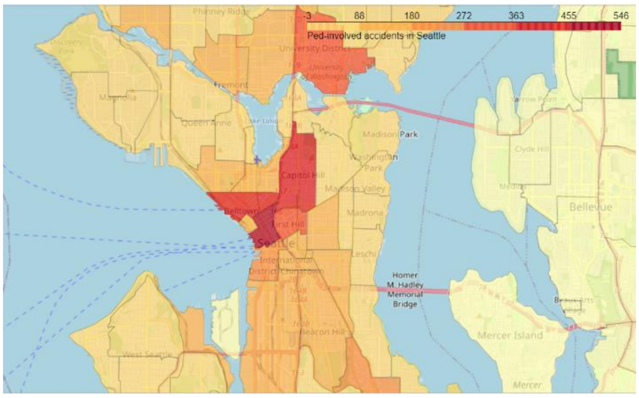

Accidents realated with Weather Conditions
Pedestrian
Cycle
57.9%
72.27%
24.08%
10.76%
14.35%
14.61%
2.62%
2.07%
0.32%
0.09%
0.32%
0.13%
0.3%
0.04%
0.09%
0.02%
0.02%
0.02%
Clear
Raining
Overcast
Unknown
Fog/Smog/Smoke
Other
Snowing
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt

Finally, we check the areas with major affection of accidents and areas regarding weather and road conditions to check if there is any area of major influence.



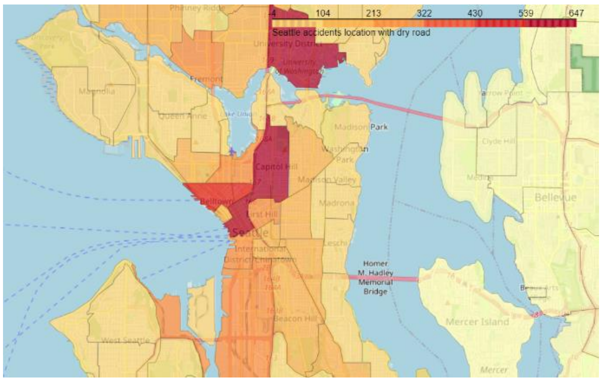
We see that there are 4 areas of high influence in accidents, which are the most transited in Seattle as are the city center and Business center. To identify any other root causes we are going to continue deeper the analysis introducing some features.

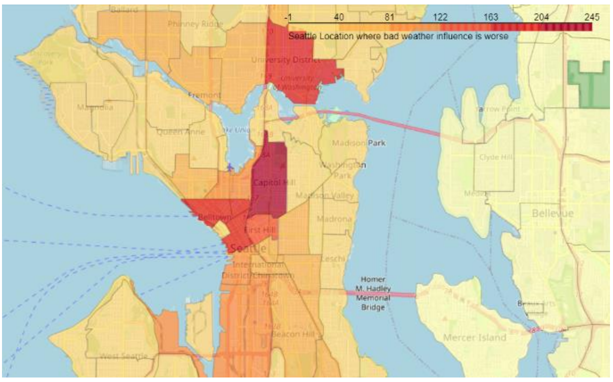




Comparing cycle-involved versus pedestrian-involved accidents, we spot that cycle-involved accidents mainly take place in two neighborhoods (one of them the University District), when pedestrian mainly take place in Central Business District. We can determine that in the university district, where the use of bicycles is more extended, we should check if there are enough facilities to move around in that means of transport. On the other hand, in the Business District, since most of the accidents are during the day and there is no data on driver inattention, the most likely cause is stress or pedestrian rush, the measures to be taken could be to check the level of pedestrianization of certain areas or the burying of the accesses to the buildings, although they require strong capital investments.

By checking the influence of the road condition or weather related to the areas of impact, influence is the same, so probably these two causes are not correlated.





4. Modelling

4.1. Pre-processing

The models aim is to predict the severity of an accident, considering that the variables of Severity Code only takes two values in our database: 1 for Property Damage Only and 2 for Injury Collision, it is not necessary to encode this variable. Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and under the influence. For lighting condition: Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was assigned 0, Mushy was assigned 1 and Wet was given 2. As for Weather Condition: 0 is Clear, Overcast is 1, Windy is 2 and Rain and Snow was given 3. For Junction Type: 0 is intersection related, 1 is not intersection related, 2 is Driveway Junction and 3 is ramp junction. As for collision type: 0 is parked car or intersections, 1 is pedestrian 2 is cycles, 3 is others and 4 is most severe accidents as sideswipe or Head-on. In general, 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either 'Other' or 'Unknown', deleting those rows entirely would have led to a lot of loss of data which is not preferred. In the other hand, rows with NaN values had been removed.

Then, it is necessary to transform dataframe to numpy arrays in order to run the classification algorithms and to split the data into train and test spaces. First one to develop a machine learning model and second one to measure its accuracy. As we have a lot of data for this analysis, the split will be 55% of data for the train set and 45% for the test set.

4.2. Machine Model Selection

The machine learning models used are k-Nearest Neighbor, Decision Tree Analysis and Logistic Regression.

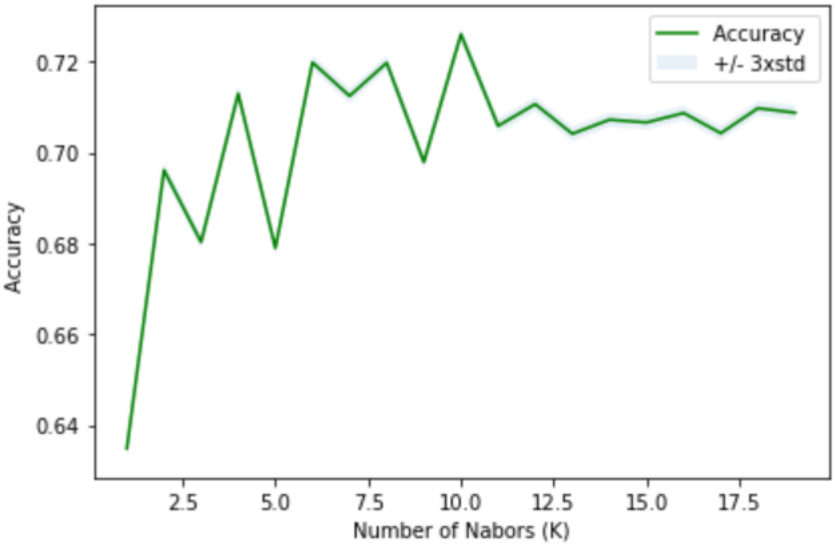
- K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance).
- Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

There is another classification model: Support Vector Machine (SVM). But it is not have been tested for this study due to its inaccuracy for large data sets, while this data set has more than 180,000 rows filled with data. Furthermore, SVM works best with dataset filled with text and images.

5. Results

5.1. K-Nearest Neighborhood

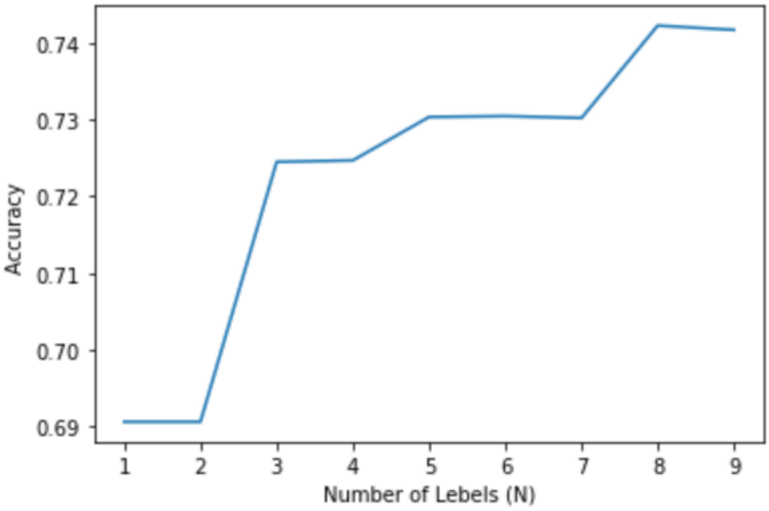
K-Nearest Neighbor classifier was used from the scikit-learn library to run the k-Nearest Neighbor machine learning classifier on the Car Accident Severity data. The best K, as shown below, for the model where the highest elbow bend exists is at 10.



5.2. Decision Tree

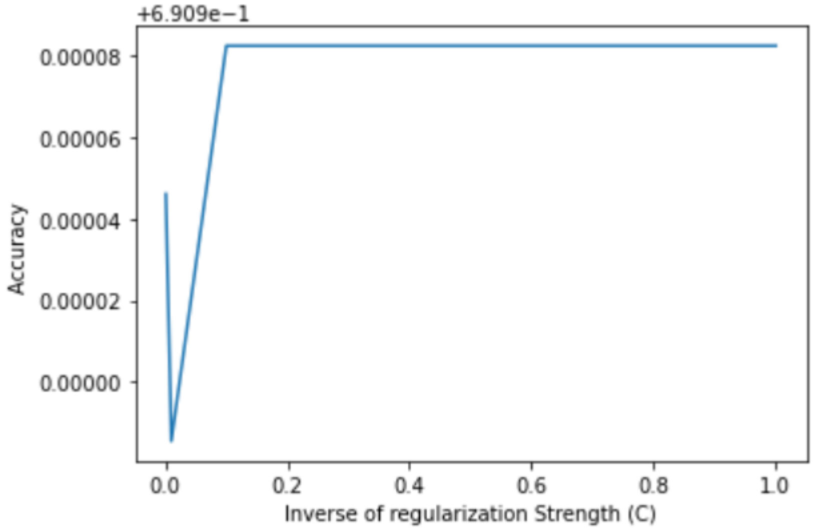
Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was 8.

```
The best accuracy was with 0.7422023802302588 with k= 8
```



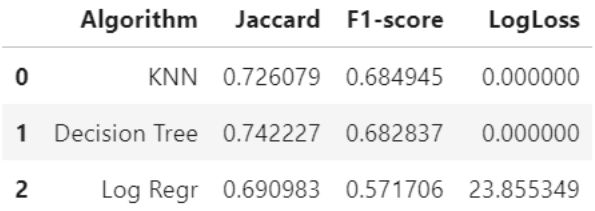
5.3. Logistic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data. The C used for regularization strength was ‘0.01’ whereas the solver used was ‘liblinear’.



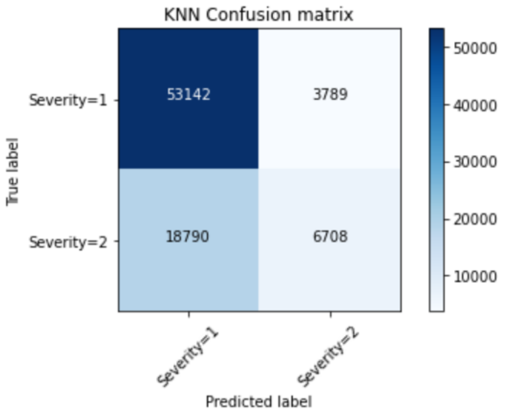
6. Discussion

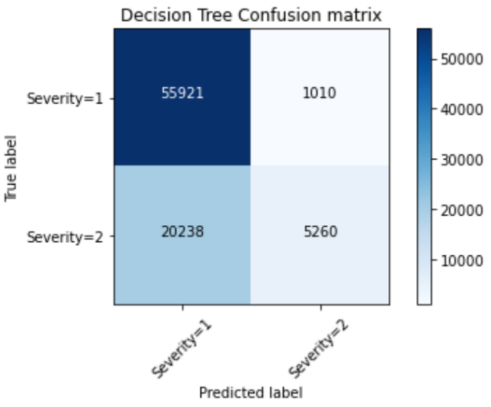
For the discussion we need to crosscheck the results predicted by the model versus test set, to compare the accuracy of the models. There are different methods to measure the accuracy, but for this study we will use: Jaccard Similarity Score, F1 Score and Logistic Loss.



	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.726079	0.684945	0.000000
1	Decision Tree	0.742227	0.682837	0.000000
2	Log Regr	0.690983	0.571706	23.855349

Clearly KNN and Decision Tree methods are the ones that give a more accurate prediction, as both parameters are quite similar, we will compare them using their Confusion Matrix:





The behavior of both algorithms is quite similar. As the dataframe was unbalance, both are better predictors for severity 1 cases than severity 2. For these case, as we prefer our solution to be conservative, it offers a better approach KNN method, as it reduces the false severity 1 amount, at the expense of working worse predicting severity 1.

7. Conclusion

By comparing both models with benchmarks within the industry, we realize that both methods (KNN and Decision Tree) work weel, but not as good as the benchmarks. These models could have performed better if a few of the following conditions were present and possible:

- A balanced dataset for the target feature
- Less missing values within the dataset for variables such as Overspeeding or Underinfluence
- More features, such as precautionary measures taken when drivin, traffic lights, etc.