Introduction

According to WHO, 1.35 million people die each year on the Word's roads, amongst which a 54% corresponds to vulnerable road users [1]. Traffic Collisions is a multidimensional issue related to a wide range of factors such as human factors (e.g. Motor vehicle speed, alcohol, distraction), road design, vehicle design and maintenance as well as sociological factors [2]. Furthermore, It is observed that minor and serious accidents are more frequent in urban areas, whereas fatal accidents are more likely in rural areas and the number of accidents in urban area depends on population size superlinearly [3].

Seattle, a large seaport city on the West Coast of the United States [4] is one of the 10 most congested cities in North America in terms of traffic and the 110th worldwide [5]. Recent reports rank Seattle as the 10th worst city to drive in the U.S. [6], fact that in accordance with Seattle's population (city population :753,675 people [7], metropolitan area population: 3,98 million [8]) might indicate a large number of traffic collisions.

The current project aims to explore data on Seattle's collisions for the time range of 2004 to 2020 and to implement prediction models to predict the different accidents' severity.

Data

The Dataset that is used by the project was generalized by the Traffic Records Group of the SDOT Traffic Management Division. It involves information related to the severity of the accident (ranked from 1 to 5) and information related to objective dimensions of the accident such as time, place as well as light, road and weather conditions.

The Dataset is provided in .csv format and involves 38 columns and 194674 rows and the time range that is covered is 16 years (01/01/2004 to 20/05/2020).

The current project uses the following data for the analysis:

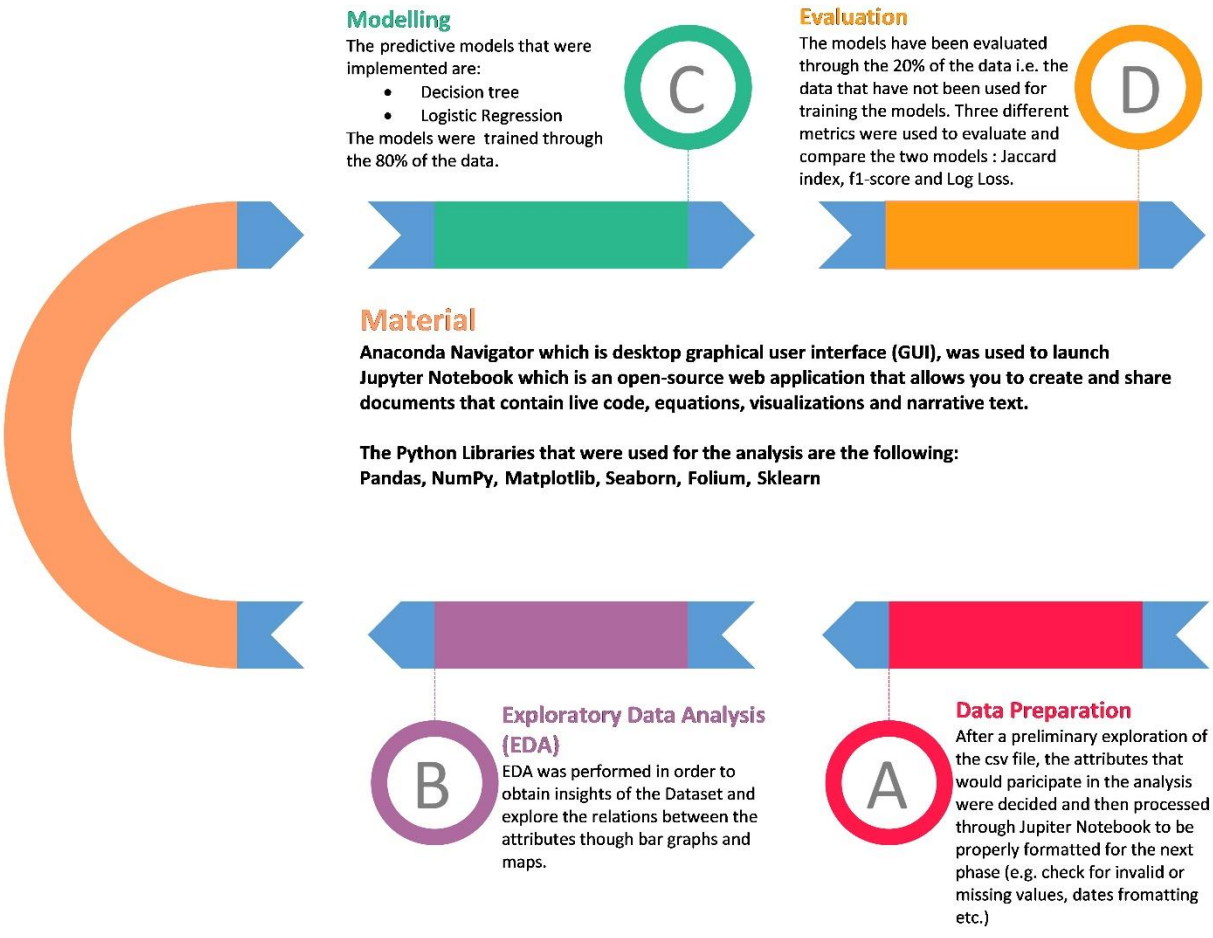
Attribute	Data type,length	Description
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
X		Longitude of the collision
Y		Latitude of the collision
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • 3—fatality • 2b—serious injury

		• **2—injury** • **1—prop damage** • 0—unknown

After a preliminary exploration of the database, it has been found that it involves only two categories for the severity column (1,2)

Methodology

At the Diagram below, we can see the methodology followed and the tools used to perform the Analysis:



The generalized code in Python can be found at the following GitHub Repository:
https://github.com/chriskourd/Coursera_Capstone

Results

Through the below graphs, the relation between each one of the five independent variables and the dependent variable (accident severity), is explored. At Chart A, we may see that the vast majority of the accidents take place during daylight, fact that is also confirmed by Chart D where the number of accidents present a gradual increase from 10:00 am and reach a peak at 17:00day pm. Graphs B and C present the road and weather conditions during the accidents respectively. The majority of the accidents take place while the weather is clear and the road is dry. Finally, at Graph E, we may see that the day of the week when most accidents occur is Friday.

Chart A

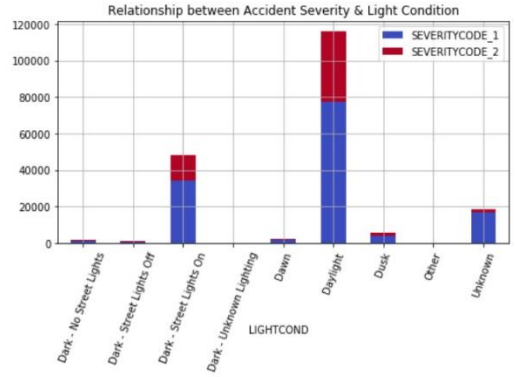


Chart B

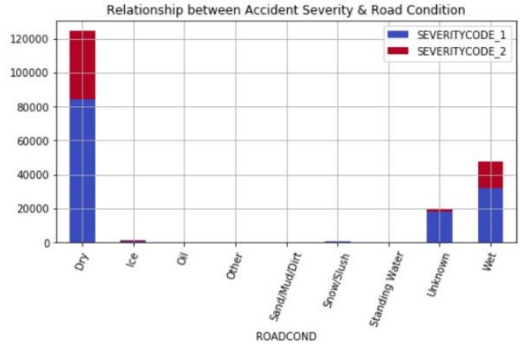


Chart C

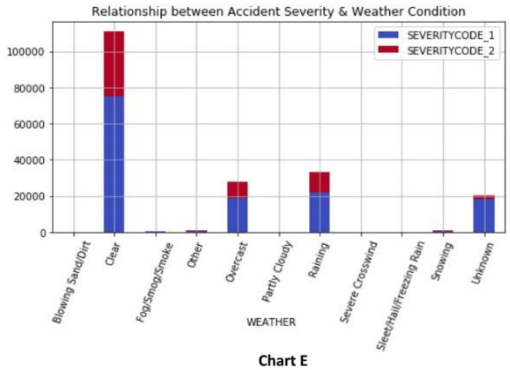


Chart D

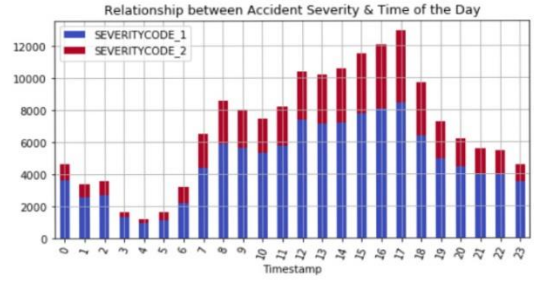
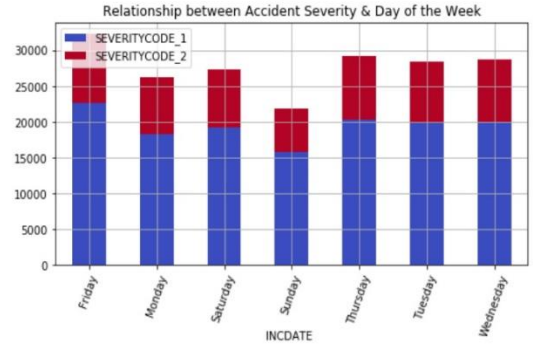


Chart E



Accuracy Report

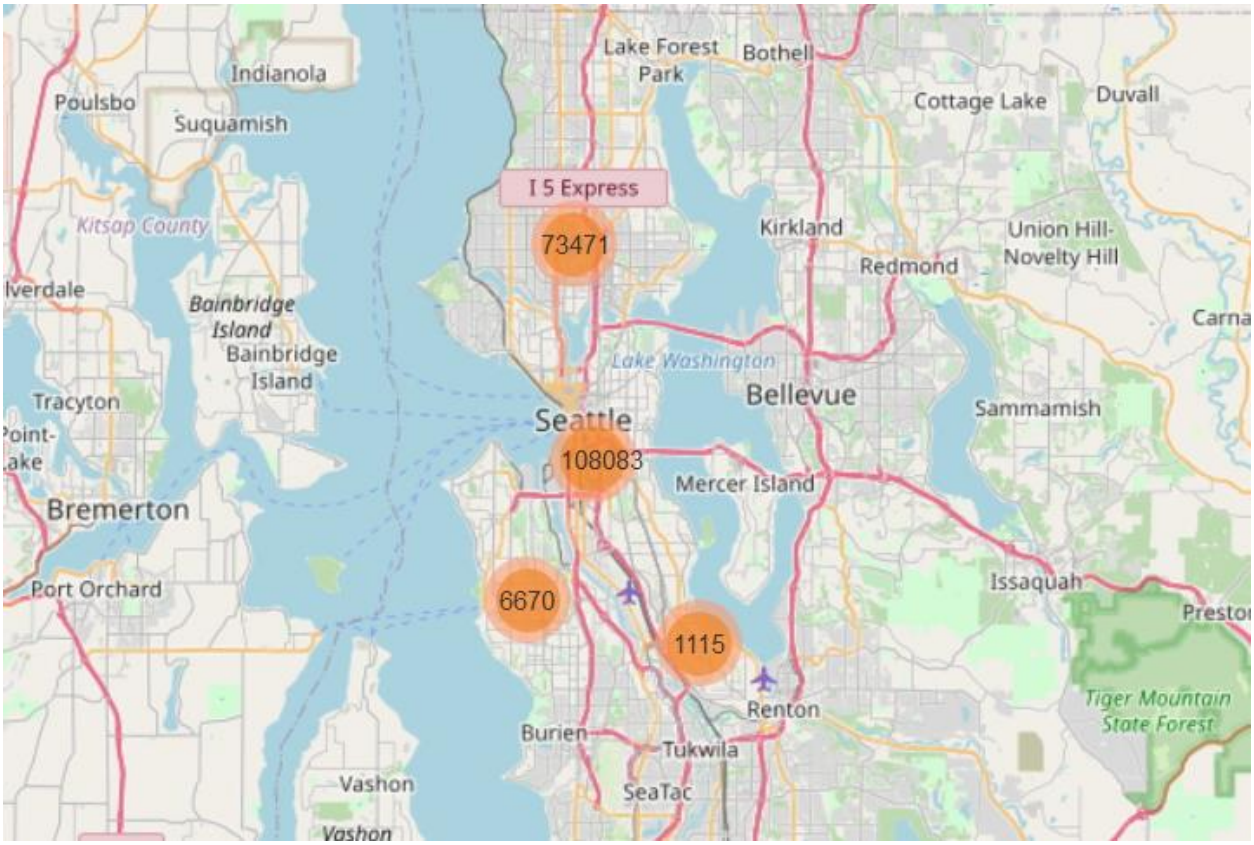
```
algorithm_list = ['Decision Tree','Logistic Regression']
Jaccard_list = [DT_jack_similarity,LR_jack_similarity]
F1_score_list = [DT_f1_score,LR_f1_score]
LogLoss_list = ['NA', LR_log_loss]

df = pd.DataFrame(list(zip(algorithm_list, Jaccard_list, F1_score_list, LogLoss_list)),
                  columns=['Algorithm','Jaccard', 'F1-score', 'LogLoss'])

df.set_index('Algorithm', inplace = True)
df
```

]:

Algorithm	Jaccard	F1-score	LogLoss
Decision Tree	0.698869	0.822746	NA
Logistic Regression	0.698839	0.822725	0.590186

Concerning the Accuracy Report, we may see that the models showed similar results for the f1-score and Jaccard Index with a slight predominance of the Decision Tree model. Jaccard similarity approaches 70% for both models and f1 score is about 0.82, which are both considered adequate scores for these metrics.

Finally, the following map shows that the majority of the car accidents (108,083) take place at the center of Seattle's metropolitan area, followed by the north (73,471), south-west (6,670) and south-east (1,115) areas.



Discussion

The results showed that both models have performed well taking into account the final metrics. The Jaccard Index indicates that the predicted values and actual values for the accident severity are similar by 70%, while an f1- score of 0.82 suggests a very good precision and recall. However, the metrics do not exactly imprint the success of a model. The f1-score index has been a subject of criticism since it gives equal importance to precision and recall [9] and the Jaccard index may be affected by the dataset size [10].

Conclusion

The current project attempted to implement prediction models which will predict the severity of Seattle's car accidents in the future. The results showed that, given the information that the used dataset provides, such a prediction is possible and reliable according to the chosen evaluation metrics.