1. Introduction

1.1 Background

Seattle, a seaport city on the West Coast of the United States, is the largest city in both the state of Washington and the Pacific Northwest region of North America. As of today's date, Seattle is home to a population of 3.4 million people. With the ever-growing population, the total number of personal vehicles had also been increasing exponentially, with the latest number at 457,000, as of 2018. Another research had also shed light on the ownership rate stating that about 81% of Seattle households owned at least 1 car. The increase in car ownership rates will no doubt lead to a higher number of accidents on the road. Statistics done by the World Health Organization showed that approximately 1.35 million people die each year as a result of road accidents, with the average of 3,700 people losing their lives every day on the road.

1.2 Problem

Road traffic accidents cost most countries an estimated 3% of their gross domestic product. The National Highway Traffic Safety Administration (NHTSA) of the USA suggests that the economical and societal harm from accidents can cost up to $871 billion a year to the US. According to the Annual United States Road Crash Statistics journal, every 16 minutes, a car accident that results in death occurs. This project aims to understand the different causes of road accidents, factors to the severity of accidents and eventually predict the severity of accidents for future analysis.

1.3 Interested Stakeholders

The analysis of the severity of accidents will be beneficial to the Public Development Authority of Seattle which works towards improving road factors. This project can also benefit the citizen as they would know how to reduce and mitigate the risk of being involved in a car accident by taking the necessary precautions. Besides that, finding out the causes and possibly predicting the severity of the accident can be useful to car, life and health insurance companies to forecast their strategies on how to produce and market their insurance products.

2. Data Understanding

2.1 Background

The dataset provided by IBM has a total of 194673 observations, with a variation in the total number of observations in each column. Specifically, the dataset consists of 38 features containing both categorical and numeric data and the features provide detailed descriptions of keys provided for each type of incident. However, the dataset had many empty columns which could have been beneficial to the project if the data is present. The columns above are referring to 'pedestrian granted way or not', 'segment lane key', 'cross walk key', and 'hit parked car'.

The machine learning model aims to predict the severity of an accident given the features attached to it. For the variable Inattention, Speeding and Under The Influence, Y is given a value of 1 whereas N and NaN is given a value of 0. For Lighting Conditions, Light is given 0, Medium is given 1 and Dark is given 2. For Road Conditions, Dry is assigned 0, Mushy is assigned 1 and Wet is assigned 2. Lastly, for Weather Conditions, Clear is dictated as 0, Overcast and Partly Cloudy is dictated as 1, Windy is dictated as 2 whereas Rain and Snow is dictated as 3. In some variables, there were unique values such as 'Other' and 'Unknown'. Deleting these rows entirely would lead to a loss of a lot of data which is not preferable.

Hence, to overcome this issue, arrays were made such as to allow each column's unknowns to be encoded based upon the original column and have an equal proportion of the elements. Then, the array was imposed onto the original column to replace the data which had 'Other' and 'Unknown'. The entire process of data cleaning led to a loss of approximately 10000 rows which had NaN data, whereas other rows mentioned before were filled accordingly.

2.2 Features Selection

After exploring the dataset, a total of 6 variables were choses to be features along with the target variable being Severity Code. These are the features to be used to uncover the insights of the dataset.

Feature variables	Descriptions
INATTENTIONIND	Is driver attentive during driving (Y/N)
UNDERINFL	Is driver driving under influence (Y/N)
WEATHER	Weather conditions during accident
ROADCOND	Road conditions during accident
LIGHTCOND	Light conditions during accident
SPEEDING	Is driver driving above the speed limit when accident occur (Y/N)

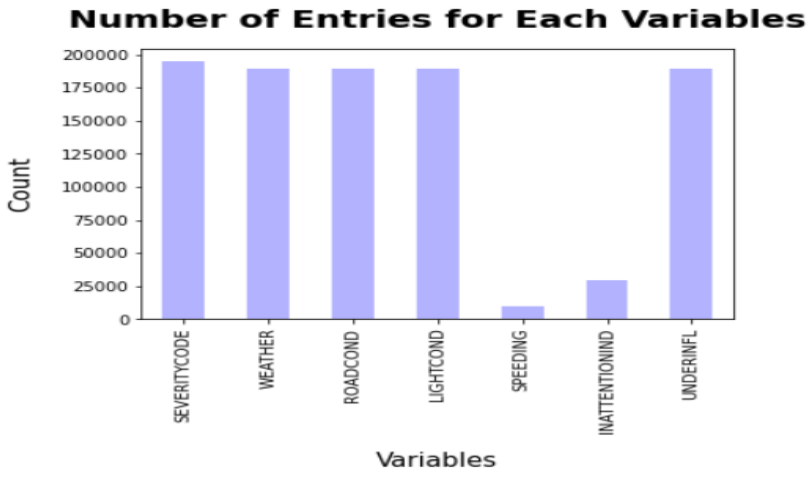
3. Methodology

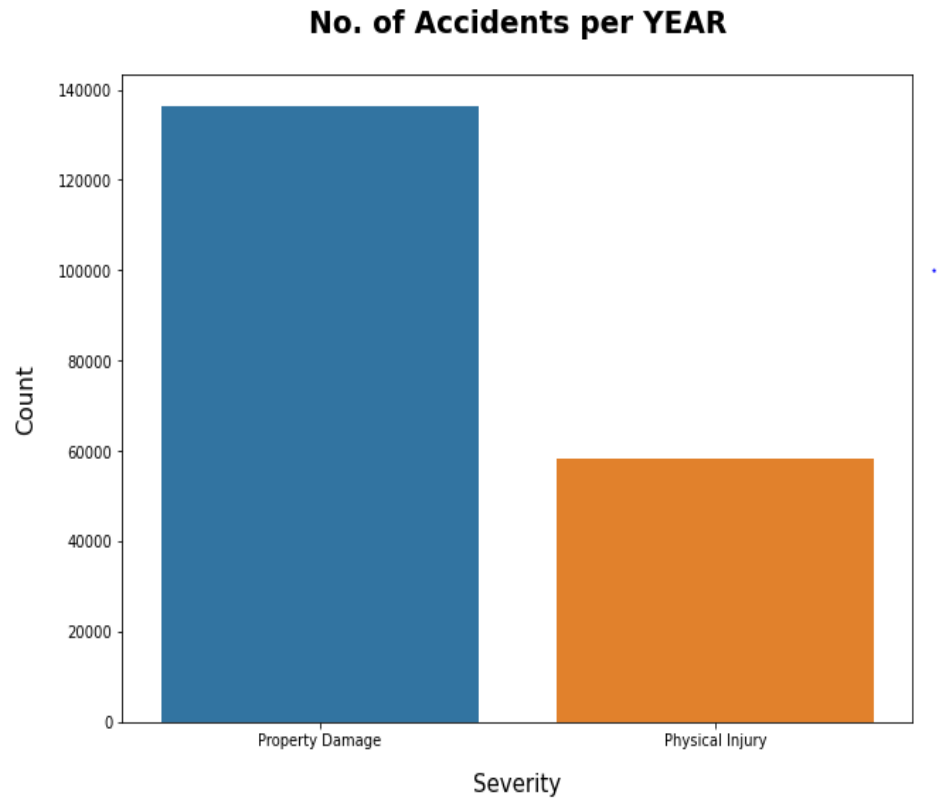
3.1 Data Collection

The dataset used for data analysis is based upon the car accidents which have taken place within the city of Seattle, Washington from the year 2004 to the year 2020.

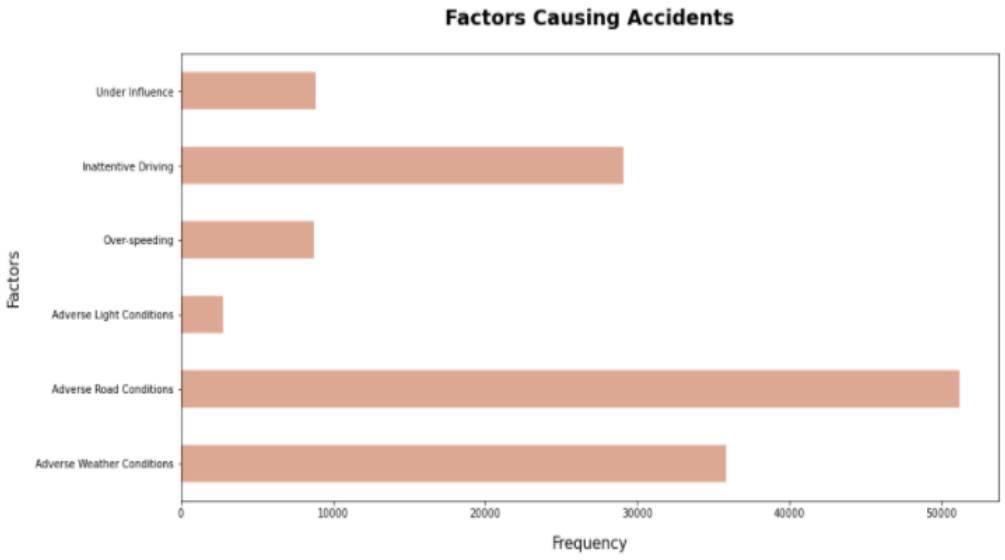
3.2 Exploratory Analysis

The features set picked and the target variables are both categorical variables with Weather, Road Conditions and Light Conditions being above level 2 categorical variables, meaning that they have more than 2 options in the categorical group. A few pictorial charts had been developed to further understand the dataset.





The figure above shows the distribution of the target variable after data cleaning had taken place. It can be seen that the dataset is supervised, however, the distribution is unbalanced in an approximate 1:2 ratio in favour of property damage. In a Machine Learning project, it is essential to have a balanced dataset to prevent biases from developing. Therefore, SMOTE was used from imblearn library to balance the target variable in equal proportion to have an unbiased classification model.



The graph above shows the frequency of accidents occuring for each feature. The factor counts are based upon the most adverse condition in each independent variable, categorised in Section 2.1. The factors which had the greatest number of accidents under adverse conditions is adverse road conditions followed by adverse weather conditions. The factors which contributed the least to the occurrence of an accident is adverse light condition.

3.3 Machine Learning Model Selection

The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbour. Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. In this case, a Multinomial Logistic Regression took place. The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time develop an associated decision tree. k-Nearest Neighbour is a simple algorithm that stores all available cases and classifies new cases based on the similarity measure (based upon Euclidean distance). Support Vector Machine (SVM) model is not used as the model is inaccurate for a large data set, in this case, the dataset used had over 180,000 rows. Additionally, SVM model works best when dataset is filled with test and images.

4. Results

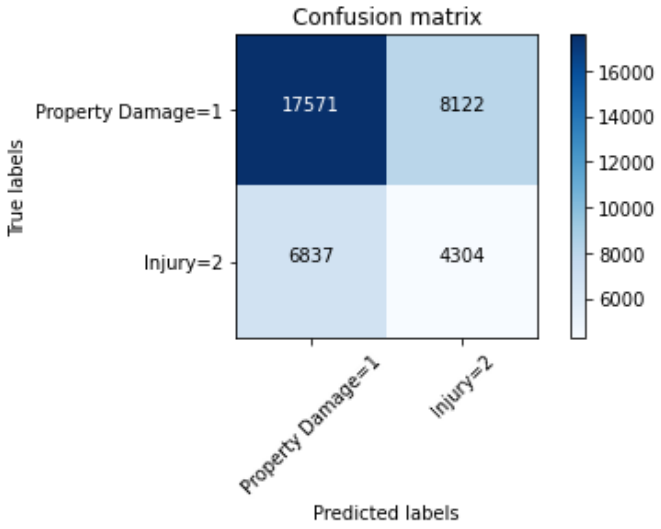
4.1 Decision Tree Analysis

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 ‘6’.

4.1.1 Classification Report

	Precision	Recall	F1-Score
1	0.72	0.68	0.70
2	0.35	0.39	0.37
Accuracy			0.59
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.60

4.1.2 Confusion Matrix



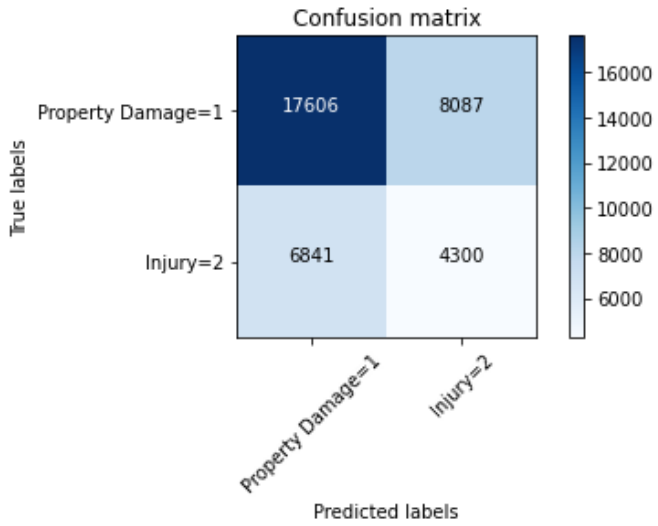
4.2 Logistic Regression

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 '0.1' whereas the solver used was 'liblinear'.

4.2.1 Classification Report

	Precision	Recall	F1-Score
1	0.72	0.69	0.70
2	0.35	0.39	0.37
Accuracy			0.59
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.60

4.2.2 Confusion Matrix



4.3 k-Nearest Neighbour

k-Nearest Neighbour from the scikit-learn library was used to run the k-Nearest Neighbour Classification model on the Car Accident Severity data. The best k was determined to be at 4 where the accuracy of the model is at its peak.

4.3.1 Classification Report

	Precision	Recall	F1-Score
1	0.70	0.88	0.78
2	0.35	0.15	0.21
Accuracy			0.66
Macro Avg	0.53	0.51	0.50
Weighted Avg	0.60	0.66	0.61

5. Discussion

Algorithm	Average F1-score	Property Damage (1) vs Injury (2)	Precision	Recall
Decision Tree	0.60	1	0.72	0.68
		2	0.35	0.39
Logistic regression	0.60	1	0.72	0.69
		2	0.35	0.39
k-Nearest Neighbour	0.61	1	0.70	0.88
		2	0.35	0.15

F1-score is a measure of the accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall are shown by the F1-score of 1, which is the highest value of the F1-score, whereas a completely imperfect model will show a F1-score of 0. The average F1-score for the 3 models are almost similar with the highest being the KNN model. Next, precision refers to the percentage of results which are relevant, in simpler terms, it can be seen as how many of the selected items from the model are relevant. It can be calculated by dividing True Positives by the sum of True Positives and False Positives. In terms of precision, the best performing model will be both Decision Tree and Logistic Regressions. Furthermore, Recall refers to the percentage of the total relevant results correctly classified by the algorithm. Among the 3 models, the recall for Property Damage and Injury is the best in Logistic regression where the outputs are balanced.

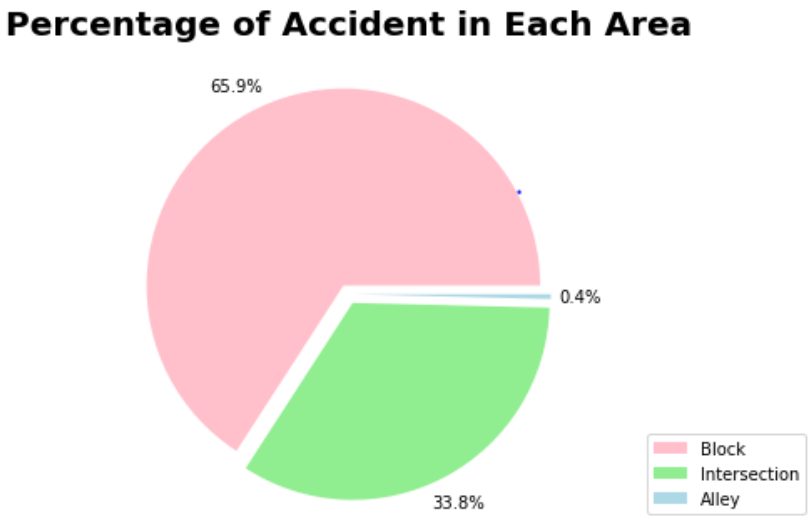
6. Conclusion

When comparing all the models performed by their F1-scores, Precision and recall, we have a clearer picture in terms of the accuracy of the 3 models individually as a whole and how well they perform for each output of the target variable. When looking at F1-scores, the best model would be the KNN model however the recall shows that the output prediction is not well balanced. Hence, among both Logistic Regression and Decision Tree, Logistic Regression would be the best model as it shows a well-balanced recall and also a good F1-score.

7. Recommendation

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders, as below.

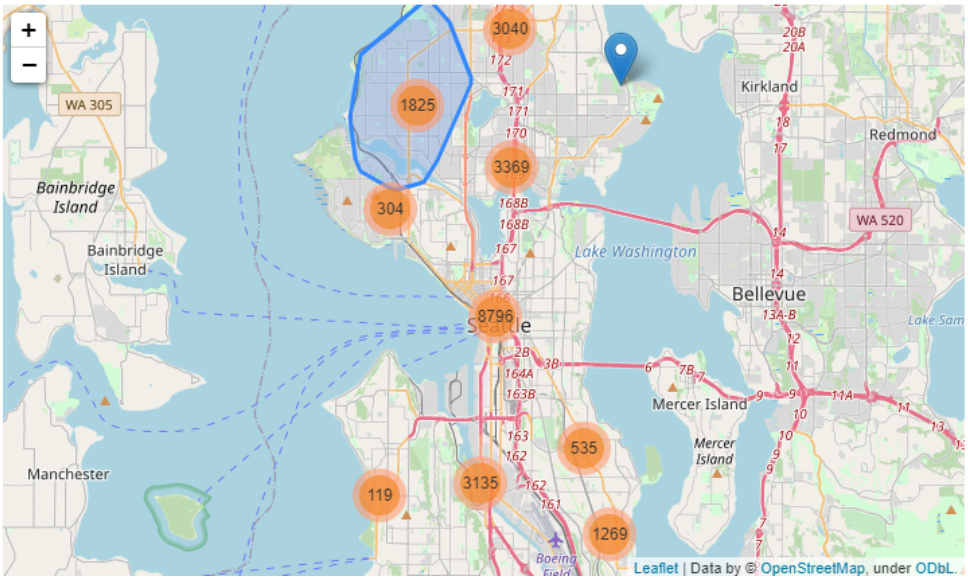
7.1 Public Development Authority of Seattle (PDAS)



PDAS can take the following measures to reduce the number of accidents:

- Increase the number of safety signs around blocks and intersections.
- Increase investment to better the road and lighting conditions for areas with higher instances of accidents recorded.

7.2 Citizen



A higher concentration of accidents can be seen on the main roads of the city, especially near the highway in the city centre. Drivers should be careful and take the following steps to reduce the chances of being involved in an accident:

- Be extra careful on the highway going through the city centre/