**1. Introduction**

**1.1 Background**

According to World Health Organization, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. Seattle is a port city in West Coast of United States, The city has found itself "bursting at the seams", with over 45,000 households spending more than half their income on housing and at least 2,800 people homeless, and with the country's **sixth-worst rush hour traffic**. Seattle has the 8th worst traffic congestion of all American cities, and is **10th among all North American cities** according to Inrix. Seattle is also referred to informally as the "Gateway to Alaska" for being the nearest major city in the contiguous U.S. to Alaska, "Rain City" **for its frequent cloudy and rainy weather, and "Jet City". Seattle recorded the highest number of car accidents in the state that year (2018), at 14,508 in Washington.**

**1.2 Problem Definition**

We are trying to create a model to predict the accident severity by road and weather conditions so that the Drivers may act accordingly to the warnings provided. They may change routes, Drive to alternate direction or drive carefully. This will help in better traffic regulation and also provide warnings in case of Severe accident (Involving fatality is predicted).

Our goal is to create a machine learning model so that we can predict that under certain given weather or road conditions how severe is the accident which has the possibility of occurring. To do this we are using the data provide by the Seattle website.

**2. Data acquisition and cleaning**

**2.1 Data Source**

The source of the data is <http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0/data>

This data set is hosted by City of Seattle at an open data platform (Also available at other open data sources). The meta-data can be obtained form <https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf>.

**A description of the data and how it will be used to solve the problem.**

The data is downloaded from the website is form 2018, and it contains 221738 rows and 40 columns. The target Variable is the accident severity which contains four severity levels namely 1, 2, 2a, 3 on an increasing level of severity.

All the feature Names are-

'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT\_COLCODE', 'SDOT\_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST\_COLCODE', 'ST\_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'.

**Classification of the features**

|  |  |
| --- | --- |
| **Locational Data** | 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC' |
| **Severity Description** | 'SEVERITYCODE', 'SEVERITYDESC', 'COLLISIONTYPE' |
| **Count of Entities Involved** | 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT' |
| **Injury Levels** | 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT' |
| **Date and Time data** | 'INCDATE', 'INCDTTM' |
| **State Designated Code and There Description** | 'SDOT\_COLCODE', 'SDOT\_COLDESC' |
| **Weather Road and Driver Conditions** | 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'HITPARKEDCAR'. |
| **State Designated Codes for Crosswalks etc.** | , 'ST\_COLCODE', 'ST\_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', |