**II. DATA**

The data that we will use to create the model, is the *Road Accident Severity Data* from the *Seattle State Department of Transport*. It is a csv file that contains 194673 entries and 36 attributes. Many of these 36 attributes are not really helpful and we will be cleaning the unnecessary data in the next stages. This file contains data about human factor’s play in these accidents and the details of damage for each instance along with dates and other ‘Yes’ and ‘No’ questions regarding whether the driver was under influence of alcohol or not, or whether the accident involved speeding or not ,or whether the pedestrian was granted the right of way to name some. The file also contains data pertaining to number of pedestrians, number of vehicles, number of bicycles and number of total people involved in each accident. This dataset has an attribute that classifies the accident into 5 different categories labelled as a number from 0 to 3. SEVERITYCODE is the attribute of interest here. A code that corresponds to the severity of the collision:

• 3- Fatality

• 2b- Serious Injury

• 2- Injury

• 1- Property Damage

• 0- Unknown

We will extract data for weather, road and light conditions, and the time of day and location into a new data-frame which will be used for our model. The columns for the same would be ROADCOND, WEATHER, LIGHTCOND, LOCATION, X, Y, INCDTTM and INCDATE. We will clean this data, remove NaN values and categories that have a very small sample size in order to balance the data. Once data extraction and cleaning are completed we will move ahead with visualizations and modelling using either KNN, Decision Tree or Logistic Regression.