**2. DATA**

The data that we are going to use for creating this project is downloaded form the link provided in the course itself. Upon inspecting the data we can sat het the sources of the data is Seattle state department of transport. This data has been provided to us in the form of comma separated values or .csv file and it contains 36 features and 194673 rows.

We will clean and remove the unnecessary features and apply the ml algorithms over that data.

This dataset has an attribute that classifies the accident into 5 different categories labelled as a number from 0 to 3. SEVERITYCODE is the target feature that we have to find about. A code that corresponds to the severity of the collision:

• 3- Fatality

• 2b- Serious Injury

• 2- Injury

• 1- Property Damage

• 0- Unknown

We will be extracting the feature data for the major features such as weather, road and light conditions, time of day and location and use this particular feature for modelling the ml problem. We will clean this data, remove null(nan) values and categories that have a very small sample size to avoid the problem of biasing appearing in the problem. Once data extracted and cleaning is completed, we will move ahead with data visualizations through which we can perform the exploratory data analysis part and future understand and processes the available data. Once the final data is present we will apply the ml algorithms of KNN, Decision Tree or Logistic Regression and finally plot the results. We will be finding the parameters of Jaccard, f1 score and log loss for all the 3 algorithms.