# A description of the problem and a discussion of the background.

## Problem

Extreme weather and unsuitable road conditions may lead to an accident. The problem is to predict the severity of an accident given the weather and the road conditions. This information can be used to change the driving to be more careful or to change the mode of travel if possible.

In this project, I attempt to predict road accidents before they happen, so the drivers can take precautions to avoid traffic jam due to accidents or even prevent accidents from happening

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

The data set used in this project came from SDOT GIS Seattle(Data Set, Meta Data). It contains the speed, light, road condition, severity, etc. for the past road accidents. The idea is to use several supervised machine learning techniques to predict the severity given the various road conditions.

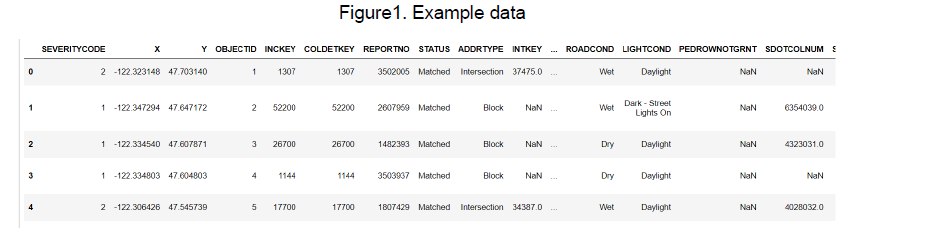
The data to be used in the Capstone project is the example dataset named "Data-Collisions.csv".

The dataset includes all collisions provided by SPD and recorded by Traffic Records. Collisions will display at the intersection or mid-block of a segment. The timeframe of the data is from 2004 to Present. The data is updated on a weekly basis. The metadata has been provided in the pdf.

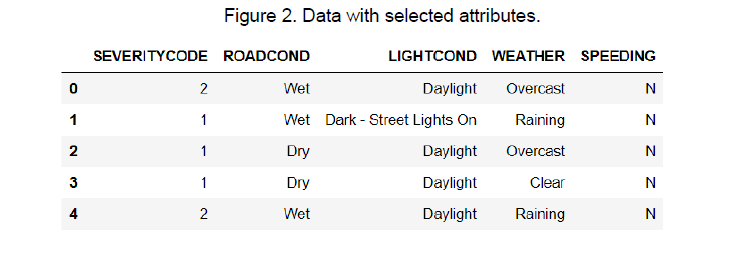
Each line in the dataset represents a single traffic accident and its various properties. The first column is the label(severity).

The remaining columns have different types of attributes which can be used to train the model. The model is supervised machine learning, so these observations are used to train and validate the machine learning model. The label for the data set is severity, which describes the fatality of an accident. The data has unbalanced labels, so steps should be done to balance the data, otherwise, it will create a biased machine learning model.

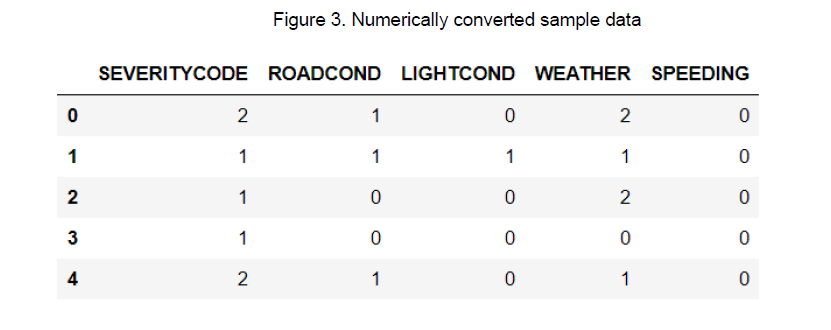
In total, there are 37 attributes/columns. Some of the attributes have missing data, so some data cleaning is required. There are both numerical and categorical types of data, such as location, number of people involved and collision types. Some or all can be used to train the model.



For the scope of this project, I'll look at how "SPEEDING", "ROADCOND", "LIGHTCOND", "WEATHER" parameters affect "SEVERITYCODE". Keep only weather in (Clear, Raining, Overcast), road condition in (Dry, Wet) and light condition in (Daylight, Dark - Street Lights On) so that the training set will not be too skewed

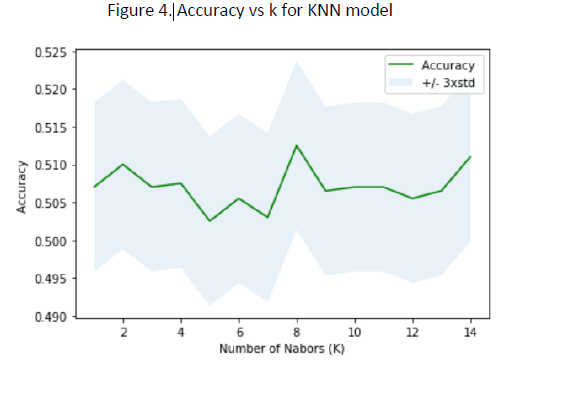


Since this is a binary classification problem (severity code= 1 or 2), I'll use K-nearest neighbors and Logistic regression techniques. KNN is chosen because its performance on dealing with a large set of data. I also chose Logistic Regression because it provides the probability for detecting accidents. To begin with, convert the attribute labels to numerical values, and then 5000 samples from each severity label were randomly picked.



## KNN

For KNN, k=8 was found to give the best accuracy = 0.5125.



## Logistic Regression

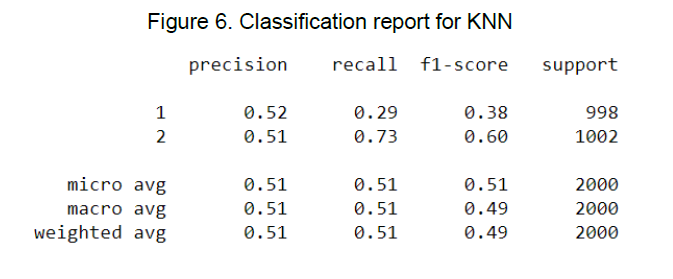
For Logistic regression, the regulation coefficient was chosen to be c=0.001, which yields a Jaccard similarity score of 0.5055

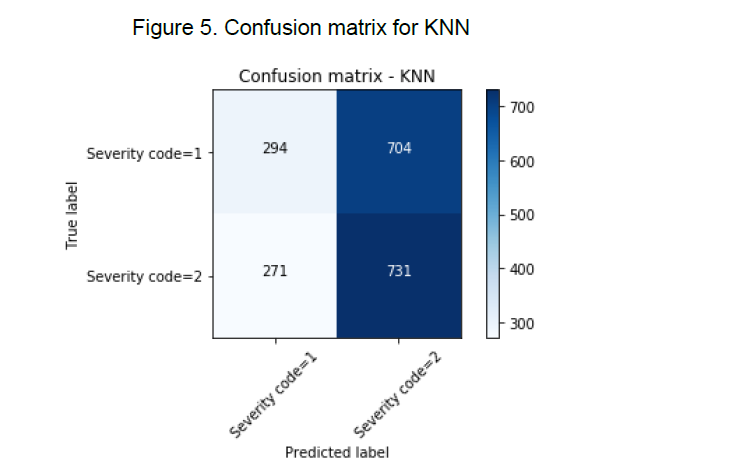
# Results

For results, confusion matrices were plotted for both KNN and LR models.

## KNN

The classification report for KNN follows.





### The below figure shows the classification report for Logistic Regression

