

Describe data

In the shared dataset, there are 37 attributes (columns). The first column tells the severity of the accident. The remaining columns have different types of attributes. Some will be used to train the model. The label for the data set is severity, which describes the fatality of an accident. Some of the shared data has unbalanced labels which may create unbiased ML model so that will require balancing the data.

The data includes all types of collisions from 2004 to present. Collisions will display at the intersection or mid-block of a segment. Not all the attributes will be used. Only a select few attributes will be used to predict and train/test the model.

The target or label column is the 'severity'. The remaining columns have different types of attributes e.g. collision type, person count, pedestrians involved, number of vehicles in the collision, weather, road condition, light condition, whether pedestrian had right of way etc.