**Data**

For this project I have chosen a dataset that consist all the details of past accidents happened in a certain place. In the data the target variable is the severity of the accidents and we have 37 independent variables. We cannot select all these variables for model creation. So data cleaning is very essential here. First we have to select the required independent attributes for the model creation. So I choose certain variables that logically can influence the road accident severity. Those variables are:

'ADDRTYPE'- Collision address type

'COLLISIONTYPE'- Collision type

,'PERSONCOUNT'-Persons involved in the accident

'VEHCOUNT'-Vehicle involved in the accident

'JUNCTIONTYPE'- Junction type

'WEATHER'-Weather condition

'ROADCOND'-Road condition

'LIGHTCOND'-Light condition

'PEDCOUNT'-Pedestrian count

The data set had some null values. To balance the dataset I removed all the records having a null value in it. Even after deleting this rows we had more than 185000 records which was enough to train a good model.

**Encoding**

Many of the attributes was categorical in our dataset so encoding this values to numerical form is a must. Since all the categorical attributes are nominal, only nominal encoding techniques are possible here. So here I used One hot encoding and mean encoding for encoding the data.