**Data Description**

Data : Data –Collisions from Applied Data Science course

Data link: [https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv](%20%20https:/s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv)

An effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

For that a proactive approach is required, that includes a collision prevention, meaning, preventing a potential unsafe road conditions from occurring in the first place. By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities. The results of analysis and modeling can be used by these Departments to take appropriate measures to reduce accident impact and thereby improve traffic safety.

In order to mitigate the impact of data size on analysis and prediction, we present a new dataset, by using labels encoding to covert the features to our desired data type. Timeframe: 2004 to Present. The data has 194674 rows and 37 columns with wide range of attributes including Location, Severity Code, Vehicle Count, Injuries, Fatalities, Junction Type, Person Count, Weather, Road Condition, collision type, address type, speeding, Collisions type etc.

Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND' and few other attributes which do not correlate and have impact on target in an regression model.

In its original form, this data is not fit for analysis. Firstly, there are many columns that we will not use for this model. secondly, most of the features are of type object, when they should be numerical type. We must use label encoding to covert the features to our desired data type, also might need to do some feature engineering to improve the predictability of your model.