COURSERA – CAPSTONE PROJECT

Car accident severity

**DATA UNDERSTANDING**

The dataset from the SDOT Traffic Management Division in Seattle presents a part of the car accidents and their characteristics detailed in 37 columns. 194 673 collisions (=rows) were recorded in the database.

The column OBJECTID will be the unique identifier of each row.

All the columns are labelled but features are not always recorded. Only the rows having a full set of data corresponding to the selected features will be used for the model. 5081 rows have no weather data, 5012 have no roadcond and 5170 have no lightcond. After cleaning, the data base will contain 189337 rows.

Inattention and speeding could have been interesting data to build the model, but those data is missing in many rows: 164868 for inattention and 185340 for speeding. Therefore, the columns must also be delated.

Features that will be extracted from the data to build the model:

Independent variables: X; Y; ADDRTYPE; LOCATION; WEATHER; ROADCON; LIGHTCOND; INCDATE; OBJECTID

Dependent variable: SEVERITYCODE

Data must be prepared by converting it to numerical type (ex: INT for integer); some rows containing missing data must be removed.

Finally, data must be balanced in order to achieve an equal representation of both SEVERITYCODES. Therefore, evaluation metrics for classifiers (like F1-score) will be used.