Data understanding:

The dataset used for this project is based on car accidents that have occurred in the city of Seattle, Washington from 2004 to 2020. The dataset contains 37 independent variables and 194,673 rows. The independent variables include “INATTENTIONIND”, “WEATHER”, “ROADCOND” and other factors that would cause the accident. The dependent variable is “SEVERITYCODE” which contains numbers of “1” and “2”, they correspond to different levels of severity of car accidents. “1” represents for “Property Damage Only” and “2” means “Physical Injury”.

In my consideration, I will drop some non-critical and indecisive attributes. The following features which I choose to remain for building model and prediction.

1. UNDERINFL which means whether or not the driver was under the influence
2. WEATHER which represents the weather condition while the collision occurs
3. ROADCOND which represents the road conditions while the collision occurs
4. LIGHTCOND which represents the light conditions while the collision occurs.

However, the existent data contains null values in some records, the data has to be preprocessed before further processing and analyzing.