

‘Other’ and ‘Unknown’ in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

### 2.3. Feature Selection

The following 5 features were selected for this study along with the target variable: Severity Code.

Feature Variable	Data type, length	Description
INATTENTIONIND	Text,1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)