

## **2. Data Description**

### **2.1. Origin of the Data**

The dataset contains 194673 observations of collisions in the city of Seattle, including all types of collisions. The dataset is provided by SPD and recorded by Traffic Records. The timeframe of the observations goes from 2004 to the present, with a weekly update. The data is organized in 37 different attributes, including among many others for example information on the location of the collision (attribute “LOCATION”), the severity of the collision (attribute “SEVERITYCODE”), or the number of vehicles involved in the collision (attribute “VEHCOUNT”). More information on the dataset can be found on the metadata sheet at <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

### **2.2. Data Cleaning**

It must be noted that the dataset contained an important number of empty columns which could have contained beneficial information. These columns included segment lane key, pedestrian granted way or not, cross walk key, and hit parked car.

The models aim was to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Injury Collision) which were encoded to the form of 0 (Property Damage Only) and 1 (Injury Collision). Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and Under the influence. For lighting condition, Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was assigned 0, Mushy was assigned 1 and Wet was given 2. As for Weather Condition, 0 is Clear, Overcast is 1, Windy is 2 and Rain and Snow was given 3. 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas there were unique values for every variable which were either ‘Other’ or ‘Unknown’, deleting those rows entirely would have led to a lot of loss of data which is not preferred.

So as to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had

‘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)