Data

The data chosen for the study consists of all collisions provided by Seattle Police Department (SPD) and recorded by Traffic Records from 2004 to present. This is a link to the dataset https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv . The data has nearly 190,000 entries and the major attributes are as follows:

- Severity code
- Location
- Junction Type
- Collision Type
- Person Count
- Vehicle Count
- Inattention while driving
- Driving under influence
- Speeding
- Weather
- Road Conditions
- Light Conditions

Target data

Of these, the Severity code was taken as the target data as it indicates whether the collision led to human injury or only property damage.

Attributes

Person Count and Vehicle Count were taken as numerical attributes to predict the target data and Speeding was taken as a Boolean.

Junction type, Collison type, Weather, Road conditions and Light conditions were all categorical attributes, but to keep the number of attributes limited to avoid overfitting and delay in processing, Collision type was taken into consideration as it showed highest accuracy post modelling.

Target	Attribute type	Attribute name
Severity Code	Numerical	Person Count
		Vehicle count
	Boolean	Speeding
	Categorical	Collision type

Data Processing

The Collision type data consisted of code numbers given by the SDOT as per the type of collision. Based on the description of the codes, they were grouped to bins having following description

1. NOT ENOUGH INFORMATION / NOT APPLICABLE
2. MOTOR VEHICLE STRUCK MOTOR VEHICLE

3. MOTOR VEHICLE STRUCK PEDALCYCLIST/PEDESTRIAN
4. MOTOR VEHICLE STRUCK OBJECT
5. DRIVERLESS VEHICLE STRUCK MOTOR VEHICLE
6. PEDALCYCLIST STRUCK MOTOR VEHICLE/ OBJECT/ PEDESTRIAN

The Speeding data was converted from text to numeric and the Collision type data was transposed and appended to the rest of the columns.

Data Transformation

The target data was converted to an array and the attributes were fit and transformed to form an array for modelling