**Data**

As exposed in the introduction section, the objective is to predict the severity degree of an accident. Given that severity is a discreet value, what we have, in terms of Data Science, is a classification problem.

The dataset used for this analysis will be the Collisions Report provided by the Seattle Police Department, that contains the information of the Traffic Records from 2004.

This dataset contains records of collisions, that include its severity degree (the value we want to predict, what we call the “label”), and some parameters that describe the accident and its circumstances. Some of these parameters can’t be known until the accident has been analysed (number of people in the car, actors involved – two cars, a car and a pedestrian, a car and a bicycle, etc. -, speed as a factor of the collision, inattention as cause of the collision, etc.) and are specific to the accident, while others are external (road, light and weather condition, etc.) and might appear in the same combination in different accidents. We will focus our attention on the latest ones and use them as the variables to classify a potential accident and provide its potential severity degree. We might add to these features other variables that might prove to be useful to improve model accuracy, as might be location, date and time.

An example of the data records that form the dataset could be the following one (a detailed description of each data field can be found on the appendices section):

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SEVERITYCODE | X | Y | OBJECTID | INCKEY | COLDETKEY | REPORTNO | STATUS |
| 2 | -1.223.659.125 | 475.173.632 | 102231 | 116343 | 116343 | 3331532 | Matched |

|  |  |  |  |
| --- | --- | --- | --- |
| ADDRTYPE | INTKEY | LOCATION | EXCEPTRSNCODE |
| Intersection | 34902 | 26TH AVE SW AND SW ROXBURY ST | NEI |

|  |  |  |  |
| --- | --- | --- | --- |
| EXCEPTRSNDESC | SEVERITYCODE | SEVERITYDESC | COLLISIONTYPE |
| Not Enough Information, or Insufficient Location Information | 2 | Injury Collision | Pedestrian |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PERSONCOUNT | PEDCOUNT | PEDCYLCOUNT | VEHCOUNT | INCDATE | INCDTTM |
| 2 | 1 | 0 | 1 | 2010/06/17 00:00:00+00 | 6/17/2010 3:50:00 PM |

|  |  |  |
| --- | --- | --- |
| JUNCTIONTYPE | SDOT\_COLCODE | SDOT\_COLDESC |
| At Intersection (intersection related) | 0 | NOT ENOUGH INFORMATION / NOT APPLICABLE |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| INATTENTIONIND | UNDERINFL | WEATHER | ROADCOND | LIGHTCOND | PEDROWNOTGRNT | SDOTCOLNUM |
| Y | N | Clear | Dry | Daylight | Y | 10168010 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SPEEDING | ST\_COLCODE | ST\_COLDESC | SEGLANEKEY | CROSSWALKKEY | HITPARKEDCAR |
|  | 1 | Vehicle turning right hits pedestrian | 0 | 525345 | N |

The data will be analysed in order to obtain the correlation degree between the severity degree of the accident and each of the available features, and also the correlation among these features, in order to choose those that might help us determine the severity of the accident and avoid those with little or no influence as well as those that might prove to be correlated among them (and thus redundant for the purpose of our analysis).

Once a simplified dataset has been obtained, different algorithms will be applied and tested, in order to find the one with the best accuracy. There will be also a fine-tuning of the parameters of the different algorithms during this process, so comparison will be done between the optimal configuration of each of the algorithms. Data biasing will be also taken into account, applying specific strategies to reduce it if necessary.