2 Data

2.1 Data source

The data can be found in the following Kaggle data set click here.

2.2 Description

The dataset that resulted from the feature selection consisted in 839,985 samples, each one describing an accident and 29 different features. These features where the following: From the characteristics dataset: lighting, localisation, type of intersection, atmospheric conditions, type of collisions, department, time and the coordinates which are described in the Kaggle dataset here. In addition, two new features were crafted, date to perform a seasonality analysis of the accident severity and weekend indicating if the accident occurred during the weekend or not. Regarding the places dataset, the selected features where: road category, traffic regime, number of traffic lanes, road pro le, road shape, surface condition, situation, school nearby and infrastructure. The users dataset was used to craft some new features: number of users: total number of people involved in the accident. pedestrians: whether there were pedestrians involved (1) or not (0). critical age: whether there were users between 17 or 31 y.o. involved in the accident. severity : maximum gravity suffered by any user involved in the accident. Unscathed or light injury (0), hospitalized wounded or death (1) The holiday dataset was used to add a last feature, labelling the accidents which occurred in a holiday.