Data

The data is divided in 5 different data sets, consisting of all the recorded accidents in France from 2005 to 2016. The characteristics data set contains information on the time, place, and type of collision, weather and lighting conditions and type of intersection where it occurred. The places data set has the road specifies such as the gradient, shape and category of the road, the traffic regime, surface conditions and infrastructure. On the user data set it can be found the place occupied by the users of the vehicle, information on the users involved in the accident, reason of traveling, severity of the accident, the use of safety equipment and information on the pedestrians. The vehicle data set contains the ow and type of vehicle, and the holiday one labels the accidents occurring in a holiday. All five data sets share the accident identification number.

An initial analysis of the data was performed for the selection of the most relevant features for this specific problem, reducing the size of the dataset and avoiding redundancy, click here. With this process the number of features was reduced from 54 to 28.

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 place 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 user 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 years. 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.