Haladó adatelemzési módszerek - Accidents in France

Júlia Jankó, Péter Mészáros, Anna Gergály

February 2024

1 Introduction

The dataset we selected for our project details the accidents that have happened on the roads of France from 2005 to 2016. The dataset contains in-depth information about the vehicles, people and road involved in the accident as well as the weather and other driving conditions. Our goal with this project is to gain insight into what factors play a significant role in roadside accidents and their severity; aiming to be able to predict accident chances and consequences.

We downloaded the dataset from Kaggle where it has been already aggregated for the years 2005-2016, however the data is originally from an official french government website, where the data stored by year. The data is organized into 5 .csv files, one of which contains holiday dates, the other four containing data pertaining to the accidents: characteristics, places, users, vehicles.

2 Description of files

2.1 Places

Contains data about the place the accident took place, the road quality and other characteristics. There are a lot categorical data describing road type, condition, accident situation, road curvature, traffic type, terrain and infrastructure. These are fairly well filled out and can be used for the creation of dummy columns.

There are a few unknown fields in this file, dealing with a PR value and the french road numbering system, which we do not have further information about.

2.2 Characteristics

Contains the characteristics of the accident. A lot of categorical columns can be found in this .csv file. The categorical data is represented in numerical coding. For most of them I would introduce dummy columns except if it has something to do with space/time/continuity that can be expressed using numbers. It has a few time columns like: Year, Month, Day. For these we could add an additional column combining these three numbers. With that we could make some continuity in the data. Express that 2012.12.31. and 2013.01.01. are very close to each other.

There are a few problematic columns here. Mostly the spacial data is incorrect or too specific. Address is very specific, it has a street name and very few data have the same address. The long-lat columns have about 230 000 rows filled out seemingly correctly out of the 900 000.

2.3 Vehicle

Contains the properties of the vehicles. Flow direction and Occupants columns have a very high 92%-99% null ratio. Flow direction's only gives information about the numbering of the postal address numbers. Occupants is only filled out when public transportation is involved in the accident.

2.4 Users

Contains information about the users (people involved in the accident). Birth year, user type (e.g. driver, passenger, pedestrian), severity are essential information and are almost never missing. Severity could serve as a main goal of the data analysis, extremities (death and no injury) are the most common. Birth may be useful when considering the driver. There is additional information about pedestrians, their location, action and whether they were accompanied at the time of the accident. Safety equipment is also noted and when used together with the severity of the accident, could reveal useful insights. The type of trip is missing in 29% of cases. Place in the vehicle is also described, but between 2004 and 2008, it was not recorded.