INFOH423 – Project: Car accident analysis in the UK Data Mining Project

Jean-philippe HUBINONT and Mahmoud SAKR

The UK government amassed traffic data from 2012 and 2014, recording nearly half a million accidents in the process. They hire your team for a great prevention mission.

1. Identify the hotspots for car accidents. These are the places where big number of accidents happen.
2. Fatal accidents are the most important to prevent following UK government. Try to understand the causes of fatal accidents and detect if patterns in the traffic condition that often lead to fatal accidents.
3. Finally, for a given traffic condition, use an appropriate algorithm to predict-and therefore allow them to prevent, if it would lead to a fatal accident.
4. Write a report that provide a comprehensive, readable (with visualization (maps, histogram, graphs, etc.)) to answer to these three questions.

The dataset you are provided with contains 35 attributes. As hints, we remind you that:

- A proper data pre-processing will be required.
- You need to use a proper validation method for the prediction part (task 3).

They will evaluate you through the report that you shall submit at the end of the project. Their evaluation is based on:

- Your ability to deal with a huge amount of data (some useful, some not).
- Your ability to answer their objectives with suitable data mining algorithms
- Your ability to use and configure (set the parameters) and validate the results of each algorithm that you would use.
- Your ability to make a right interpretation of the results for each of these algorithms.

In order to do so, please:

- Justify your pre-processing (e.g., what attributes/tuples are useful and usable ?)
- Justify the choice of each algorithm (what problem it solves and what kind of results to expect)
- Justify each parameter used in this context (sometimes it can be common sense, or expert judgment, but explain at least why you chose a certain value),
- Present the results as clearly and relevantly as possible (using charts, tables, maps, etc.).
- Use any software you want. If you use Rapidminer, be sure to have an educational license in each group. Otherwise, your dataset will be limited to only 10000 examples.

Here is the list of the 35 attributes you can find in the file accidents_2012_to_2014.csv

- Unnamed: 0
- Accident_Index (Unique ID)
- Location_Easting_OSGR (Local British coordinates x-value)
- Location_Northing_OSGR (Local British coordinates y-value)
- Longitude
- Latitude
- Police_Force
- Accident_Severity (1 = Fatal, 2 = Serious, 3 = Slight)
- Number_of_Vehicles
- Number_of_Casualties
- Date (In `dd/mm/yyyy` format)
- Day_of_Week (Numeric: 1 for Sunday, 2 for Monday, and so on.)
- Time (Time the accident was reported, in UTC+0)
- Local_Authority_(District)
- Local_Authority_(Highway)
- 1st_Road_Class (This field is only used for junctions)
- 1st_Road_Number (This field is only used for junctions)
- Road_Type (Some options are Roundabout, One Way, Dual Carriageway, Single Carriageway, Slip Road, Unknown)
- Speed_limit
- Junction_Detail (Some options are Crossroads, Roundabouts, Private Roads, Not a Junction.)
- Junction_Control (A person, a type of sign, automated, etc.)
- 2nd_Road_Class (This field is only used for junctions)
- 2nd_Road_Number (This field is only used for junctions)
- Pedestrian_Crossing-Human_Control (Was there a human controller and what type?)
- Pedestrian_Crossing-Physical_Facilities (Was it a zebra crossing, or bridge, or another type?)
- Light_Conditions (Day, night, street lights or not.)
- Weather_Conditions (Wind, rain, snow, fog)
- Road_Surface_Conditions (Wet, snow, ice, flood)
- Special_Conditions_at_Site (Was anything broken or defective, e.g. an obscured sign?)
- Carriageway_Hazards (Was something in the way, e.g. a pedestrian, another accident, something in the road?)
- Urban_or_Rural_Area
- Did_Police_Officer_Attend_Scene_of_Accident
- LSOA_of_Accident_Location
- Year