# DATA DESCRIPTION

The file roadAccident.xlsx contains data about road accident recorded from 2000 to 2005 in UK. There are 10 variables (attributes) involved, and 20322 records. The table below shows the metadata about the dataset.

**Table**: Variables and Values for Road Accident Data

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Name | Value | Code description |  |
| ACCYEAR |  | Accident Year |
|  | 2000 | Year 2000 |
|  | 2001 | Year 2001 |
|  | 2002 | Year 2002 |
|  | 2003 | Year 2003 |
|  | 2004 | Year 2004 |
|  | 2005 | Year 2005 |
| RD\_CLS |  | Road Class |
|  | 1 | Motorway |
|  | 2 | A(M) |
|  | 3 | A |
|  | 4 | B |
|  | 5 | C |
|  | 6 | Unclassified |
| SP\_LIM |  | Speed Limit |
|  | numeric | Miles/hour, e.g., 30, 60 |
| JUNC\_DET |  | Junction Detail |
|  | 0 | Not at junction or within 20 metres |
|  | 1 | Roundabout | |
|  | 2 | Mini-roundabout | |
|  | 3 | T, Y or staggered road | |
|  | 5 | Slip road | |
|  | 6 | Crossroads | |
|  | 7 | Multiple junction | |
|  | 8 | Private drive or entrance | |
|  | 9 | Other junction | |
| LIGHT\_COND |  | Light Condition | |
|  | 1 | Daylight - lights present | |
|  | 2 | Daylight – no lighting | |
|  | 3 | Daylight – lighting unknown | |
|  | 4 | Darkness – lights lit | |
|  | 5 | Darkness – lights unlit | |
|  | 6 | Darkness – no lighting | |
|  | 7 | Darkness – lighting unknown | |

|  |  |  |
| --- | --- | --- |
| WEATH\_COND |  | Weather Condition |
|  | 1 | Fine no high winds |
|  | 2 | Raining no high winds |
|  | 3 | Snowing no high winds |
|  | 4 | Fine + high winds |
|  | 5 | Raining + high winds |
|  | 6 | Snowing + high winds |
|  | 7 | Fog or mist |
|  | 8 | Other |
|  | 9 | Unknown |
| CASU\_CLS |  | Casualty Class |
|  | 1 | Driver or rider |
|  | 2 | Passenger |
|  | 3 | Pedestrian |
| SEX\_CASU |  | Sex of Casualty |
|  | 1 | Male |
|  | 2 | Female |
| AGE\_CASU |  | Age of Casualty |
|  | numeric | e.g., 18, 25 |
| SEVE\_CASU |  | Severity of Casualty |
|  | 1 | Fatal |
|  | 2 | Serious |

Like much of the data that companies store in data warehouses, this is genuinely historical data recorded by government, and much of the interest lies in trying to discover patterns within it.

**Task to do:**

You are asked to use OpenRefine, Weka and Tableau to conduct an exploratory data mining of this data, and to produce a SHORT report about what you discover from the data.

1. Prepare and clean the data for analysing. At this stage, you are expected to undertake at least the following procedures: Understand Data and Prepare Data. At the preparation stage, you should clean, and convert the data from the XLSX format into the format that can be accepted by software tools suggested. This might include transforming data from one type to another in order to use some particular algorithms or visualisation techniques. For example, you might also need to format the data to be able to understand it in Tableau, e.g interpreting what the coded values mean; or transform from numbers to nominal values in order to use Apriori to analyse it. Therefore, you are expected to prepare more than one version of data for the analysing.

2.

* 1. Analyse the data with appropriate techniques/algorithms such as Classification, Regression, Association and Clustering algorithms. At this stage, you should find out some interesting patterns, such as under which condition it is likely to have fatal accident, which type of accident would be likely to happen on which type of road, does driving speed have impact on accident, the number of which kind of accident has decreased from year 2000 to year 2005 in the data, by using appropriate algorithms. These patterns should be represented by rules, supported by statistical information, such as accuracy and coverage.
  2. Use Tableau to visualise and explore the data set and discover interesting patterns and features in the data. Your tasks are:

1. Before beginning the analysis, formulate an initial question that would be interesting to ask of the data. For example, you might ask “*what is the relationship between weather conditions and the severity of accidents on different types of road*?” This question should be included in your report.
2. Use Tableau to gain a preliminary understanding of the data

e.g. check the distributions of each attribute and look for correlations/dependencies between pairs of attributes. You should include in the report:

1. Which layouts and visualisation encodings you used to carry out this stage of analysis.
2. Any interesting findings made.
3. Any revisions you make to your initial question in light of this step.
4. Sketch at least three possible visualisation solutions that could potentially be used to answer your (revised) question. Be sure to annotate your sketches with any interaction techniques required. Select one of these visual solutions to implement in Tableau. Include in your report:
   1. A copy of your sketches (as an appendix).
   2. A brief discussion of the layouts, encodings, and interaction techniques, considered in your sketches.
   3. A discussion of your reasons for selecting your chosen visual solution over the others.
5. Implement your chosen visual solution using Tableau and use it to answer your question. Include in your report a discussion of any interesting discoveries that you were able to make and how your visual solution helped you to make these discoveries. Include screenshots of the visualisation as evidence of your findings

3. Based on your analysis in 2, summarise the overall findings in the data. Discuss and compare the algorithms and visualisation techniques used and draw an overall conclusion with justification about which techniques are most effective for making discoveries and gaining insights into the data.

Beware of the fact that some of the classifiers only accept nominal (qualitative) attributes, the results need some interpretation in the light of common sense and basic knowledge of what the data is actually about. Also remember that the coverage and accuracy of rules generated by each algorithm, if they are available, are important.

**Report to produce at the end :**

A short report required explaining how and what you did in your experiment (enough to let an experienced user re-create what you get, e.g., clearly describe the errors you identified and the corrections made accordingly when you were processing data cleaning; the way you perform any data transformation at the data preparation stage; at the analysing stage, mention the attributes involved in each play, mention any change to parameters etc.) and will interpret and discuss the results generated. Any screenshots that are necessary can be put into an appendix which is not included in the eight page limit.

* The report, cleaned, and formatted datasets, including all datasets that are ready for the analysing tools . just zip them into one file

**Hand in:** A hard copy of the report should be handed to the Computing Programmes Office (Room C34 Merchiston), with a coursework coversheet. Matriculation card is required when you hand in your work.