SET11122 Data Analytics 2017/18

Coursework Assessment I

DATA DESCRIPTION

The file roadAccident-with errors.xlsx contains data about road accident recorded from 2000 to 2005 in UK. There are 10 variables (attributes) involved, and 20220 records (instances). 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. For example, under which conditions, an accidence would likely be a Fatal one. Unfortunately, there are a number of errors in the dataset. Before any kind of analysing, the dataset has to be cleansed fist.

The data for this coursework is available at the module’s Moodle site.

YOUR TASK

You are asked to use OpenRefine and Weka to prepare the data for analysing and to produce a SHORT report to describe how the task is done.

Task and mark allocations are as follows

1. Understand and clean the data for analysing. At this stage, you are expected to undertake at least the following procedures: Understand Data and Clean Data. In data understanding, you are expected to understand the data clearly based on the provided metadata. In data cleaning, you should be able to identify all possible errors in the dataset and correct them. **[12%]**

2. Convert the data for analysing by Weka. For this task, you are expected to convert the cleansed dataset (generated in task 1) from the XLSX format into the format that can be accepted by Weka, the .ARFF format. This might include transforming data from one type to another in order to use some particular algorithms. For example, you might need to transform

some attribute values from numeric values to nominal values in order to use algorithms that do not accept numeric values, such as Apriori. Also you might need to transform some attribute values from nominal values to numeric values in order to use algorithms that do not accept nominal values, such as k-means. Therefore, you are expected to prepare three datasets, one without any transformation, one with all nominal values, and one with all numeric values.

(8%)

Total [20%]

THE REPORT

Your report should be no longer than **FIVE** pages of A4, using at least a 12pt font. It should include a clear description of the ways you do the data cleaning, and the ways you perform any data transformation. Errors identified and corrected accordingly should be included, in a table format (suggested). Any screenshots and tables that are necessary can be put into an appendix which is not included in the five page limit.

Note

This coursework contributes 20% to the overall module assessment.

Collaboration and Plagiarism

This is an individual assessment. The work submitted should be entirely your own and will be checked against all other submissions by TurnitinUK.

Deliverables :-

- The report, cleaned, and formatted datasets, including all three datasets that are ready for the analysing tool, Weka should be zipped into a file called set11122cw1_<your matric number> and uploaded to Moodle by the submission deadline, as per instructions. For example, if your matriculation number is 40123456, your zipped file should be named set11122cw1_40123456.

Deadline: 17:00, 28 June 2018