Enterprise Database Technologies
B.Sc. in IT Management
B.Sc in Computing
Higher Diploma in Computing

CA 1 – In Groups of Two
This CA is worth 25%
Released: 13th February 2018
Upload by 9th March 2018

This CA is designed to assess your skills and abilities to pre-process the data from a data understanding and exploratory data analysis perspective in the context of the data mining process. There are two sections to the report namely **1) Getting to know the Dataset** (exploratory data analysis) and **2) Cleaning and Transforming the Dataset** (Data Pre-processing)

See moodle for details on what dataset each group must use.

While this is a group assignment, individual marks will be given if it clear that similar contributions were not provided by each student. Each task a student provided a significant contribution to, must be clearly initialled in the document.

A brief viva/presentation of findings may be required as part of this work. Any work not directly your own must be referenced using Harvard referencing.

NB: Institute plagiarism rules apply to this and every CA. These rules will be strictly enforced.

The assignment will be checked through the Turnitin Plagiarism Prevention system,

It is critical that all terms are properly explained and that the significance of your observations and comments are properly communicated to the reader.

It is expected that authors will seek to tie in appropriate material from the lectures, and readings into this report.

Marks will be awarded for completeness and correctness of the analysis, synthesis of ideas, conciseness and clarity of thought and argument.

NB The report should be no longer than 5 written pages (excluding appendices). Evidence of the use of R should be correctly referenced in the appendix. Significant graphs and statistics to justify your findings and argument should only appear in the main body of the report. Only relevant secondary information can appear in the appendix.

You are required to upload one PDF file containing your report.

A representative sample dataset is provided. It is expected that you **will use R to explore the data**. You may use **Weka** for certain elements as indicated in the assignment. The approach to the problem follows the CRISP-DM process: data understanding (also called exploratory data analysis) and data pre-processing stages.

Before starting your analysis, please familiarise yourself with the dataset provided. You can assume the sample was randomly selected from a population that is normally distributed. Please refer to the relevant documentation for the description of the predictor variables and response (target) variable.

As part of an appendix to your report provide clearly commented supporting R (where appropriate) code you used to carry out your tasks).

1. Getting to know the Dataset using R:

1. For each predictor variable, where appropriate, find the following information:

- **a.** The attribute type, e.g. nominal, ordinal, numeric.
- **b.** Percentage of missing values in the data.
- **c.** Max, min, mean, mode, median standard deviation.
- **d.** The type of distribution that the numeric attributes seem to follow (e.g. normal).
- **e.** Whether the numeric data is skewed and the type of skewness.
- **f.** The level of correlation among predictor variables. Should there be any action taken? What is the correct action?

Provide a short commentary on overall findings\observations. Discuss any preliminary evidence of any problems in the data.

2. Construct a histogram for each numerical variable, with an overlay of the target variable.

- **a.** Discuss the relationship, if any, of each of the numerical predictor variables has with the target variable (response variable).
- **b.** Which variables would you expect to make a significant appearance in any machine learning classification model? Justify your answer.

3. Construct a bar chart for each categorical variable, with an overlay of the target variable.

- **a.** Discuss the relationship, if any, each of the categorical predictor variables has with the target variable (response variable).
- **b.** Which variables would you expect to make a significant appearance in any machine learning classification model? Justify your answer.

4. Using graphical methods are there outliers (extreme values) in the data? Confirm your assertions using at least two statistical methods you are familiar with. Provide the merits\demerits of each method.

5. Investigate whether there are any correlated variables. Using R to visualize 2D-scatter plots for each pair of numeric attributes:

- **a.** Does any pair of variable look to be correlated?
- **b.** Verify your assertions using statistical methods.
- **c.** Which attributes seem to be the most/least linked to the target variable? Summarise in a table your findings concerning the predictive value of each attribute with respect to the target variable.

2. Cleaning and Transforming the Dataset – You may use R or Weka for this part.

6. Choose ONE numeric predictor variable and bin (discretise) the data using "equal width binning". Carry out the same process using k- means clustering machine learning (ML) algorithm. Try it for different values of k. Examine the bin membership for each method. Discuss your findings. Propose the optimal solution explaining why.

   [You can use Weka or R to carry out clustering. Check out the kmeans function found in the R stats package]

7. Choose a numeric predictor variable you have determined as skewed. Transform the data using the following transformation methods in an attempt to achieve normality. Comment on your findings.
   - **a.** Z-Score Standardisation
   - **b.** Natural Log Transformation
   - **c.** Square Root Transformation
   - **d.** Inverse Square Root Transformation

8. Regarding missing values, take ONE of your categorical variables and develop an appropriate Machine Learning (ML) model to impute the missing values for that attribute. Impute the values accordingly. Briefly discuss the approach you took explaining how your imputation works. [NB – you will use a supervised method with the missing values variable as the target variable]

   It has been agreed that remaining missing categorical values (if any) should be imputed using the mode for that predictor variable. Missing numeric values should be imputed using the median value for the variable. Comment on your findings as well as the actions you carried out.

9. Are there any variables which can be eliminated? Justify your answer and express the possible benefits of doing so (if any). Use three different techniques to determine your decision. *You may use Weka Data Mining tool to carry out this task.*

APPENDIX

Dataset 1 - Banking Data

The data is related to direct marketing campaigns of a major banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.1827 instances with 18 predictor variables and a binary target variable.

1- age	Age
2 - job	type of job
3 - marital	marital status (note: 'divorced' means divorced or widowed)
4 - education	Educational Attainment
5 – housing	has housing loan?
6 – loan	has personal loan? (# related with the last contact of the current campaign)
7 – contact	contact communication type
8 – month	last contact month of year
9 - day_of_week	last contact day of the week
10 - duration	last contact duration, in seconds
11 - campaign	number of contacts performed during this campaign and for this client
12 – pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously)
13 – previous	number of contacts performed before this campaign and for this client
14 – poutcome	outcome of the previous marketing campaign
15 - cons.price.idx	consumer price index - monthly indicator
16 - cons.conf.idx	consumer confidence index - monthly indicator
17 - euribor3m	euribor 3 month rate - daily indicator
18 - nr.employed	number of employees - quarterly indicator
19 - y	Output variable (desired target): has the client subscribed a term deposit? (binary: 'yes','no')

Dataset 2 - The Cardiology Patient Dataset

The dataset consists of 310 instances. There are instances that hold information about patients who are free of heart disease (healthy). The remaining instances contain information about patients who have had a heart attack (sick). The dataset contains 14 numeric attributes (some are flags, categorical and continuous) and a sixteenth attribute indicating whether the patient has a heart condition.

Attribute Information:

-- 1. (age)	Age in years
-- 2. (sex)	Sex (1 = male; 0 = female)
-- 3. (cp)	Cp: chest pain type -- Value 1: typical angina -- Value 2: abnormal angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
-- 4. (trestbps)	Resting blood pressure upon hospital admission (in mm Hg on admission to the hospital)
-- 5. (cholestrol)	Serum cholestoral in mg/dl
-- 6. (Fasting blood sugar)	Is fasting blood sugar > 120 mg/dl? (1 = true; 0 = false)
-- 7. (restecg)	Resting electrocardiographic results -- Value 0: normal -- Value 1: abnormal (i.e. having ST-T wave -- Abnormality - T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: Hyp showing probable or definite left ventricular hypertrophy by Estes' criteria
--8. (diastbpexerc)	Diastolic blood pressure during exercise. The diastolic blood pressure number indicates the pressure in the arteries when the heart rests between beats.
-- 9. (thalach)	Maximum heart rate achieved
-- 10. (exang)	Exercise induced angina (1 = yes; 0 = no) i.e. Does the patient experience angina as a result of exercise?
-- 11. (oldpeak)	ST depression induced by exercise relative to rest
-- 12. (slope)	The slope of the peak exercise ST segment -- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
-- 13. (ca)	Number of major vessels (0-3) coloured by flourosopy
-- 14. (thal)	3 = normal; 6 = fixed defect; 7 = reversable defect
-- 15. (class) (the predicted attribute i.e. the response variable)	diagnosis of heart disease (angiographic disease status) -- Value 0: Sick -- Value 1: Healthy