Comp3027J Data Mining and Machine Learning

Module Title:	Data Mining and Machine Learning
Assignment Type:	Report and code
Project Title:	Data Mining and Machine Learning Assignment Part 1
Project Date:	13/03/2019
Assignment Compiler:	Dr. Catherine Mooney
Weighting:	N/A
Due Date:	Friday, March 29, 2019, 4.30 pm
Method of Submission:	Printed report handed into BDIC academic affairs office and zip file of report (pdf) plus code (.r file) to Moodle

Task: Using the Pima Indians Diabetes Database (see last page for details) you will:

1. Design an Analytics Base Table (See lecture 2)

2. Create a data quality report (See lecture 2)
Identify and handle any data quality issues (e.g. missing values, irregular cardinality or outliers. Note: some of the variables have 0 values but it is not possible for someone's BMI or blood-pressure be 0.What are you going to do about this? See lecture 2)

3. Feature Selection
Visualize the relationships between the features (see lecture 3)
Are the features predictive, interacting, redundant or irrelevant? (see lecture 5)

4. Data preparation (normalization, binning, sampling, etc. see lectures 2, 3 and 4)
Note: There are nearly double the number of observations with class 0 than there are with class 1. Should you be concerned about this?

Deliverables

1. A printed report handed into the BDIC academic affairs office before the deadline. This should be a clearly written report detailing how you carried out each of the task 1 – 4 above and showing the results that you got. It should be written in:
    - Font: Times New Roman
    - Font size: 12
    - 1.5 line spacing
    - Use clear headings for each task (1 – 4)
    - Include tables and figures as appropriate (Use captions i.e. Table 1 followed by a description of Table 1, and then you can refer to Table 1 in you text.)
    - There is no page limit, use as much space as you need to clearly report your findings but please keep the report short and to the point. There are no marks for filling space with waffle!
    - Your report should be presented as professionally as possible. Spelling, grammar and formatting all count.
    - Include a coversheet with your name, email, UCD and BJUT student

numbers and the title “Comp3027J Data Mining and Machine Learning Assignment”

2. A zip file submitted to the moodle before the deadline. The zip file should be made up of:
   1. a commented R file named “ml-assignment.r”. This should have all the code you used to perform the tasks 1 – 4 above. Please keep the code clear and simple. Use the examples from the labs.
   2. A pdf of your report

Plagiarism

Plagiarism is a serious academic offense

I will be proactive in looking for possible plagiarism in all submitted work

Suspected plagiarism will be reported and subject to investigation

1st offense: usually 0 or NG in the affected components
2nd offense: referred to the University disciplinary committee

Students who enable plagiarism are equally responsible

http://www.ucd.ie/registry/academicsecretariat/docs/plagiarism_po.pdf

http://www.ucd.ie/registry/academicsecretariat/docs/student_code.pdf

http://libguides.ucd.ie/academicintegrity

Pima Indians Diabetes Database

- Sources: (a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases (b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader Applied Physics Laboratory The Johns Hopkins University Johns Hopkins Road Laurel, MD 20707 (301) 953-6231 (c) Date received: 9 May 1990
- Relevant Information: Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.
- Number of Instances: 768
- Number of Attributes: 8 plus class
- For Each Attribute: (all numeric-valued)

  Pregnancies: Number of times pregnant
  Glucose: Plasma glucose concentration

BloodPressure: Diastolic blood pressure
SkinThickness: Triceps skin fold thickness
Insulin: 2-Hour serum insulin
BMI: Body mass index
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age
Outcome: Class variable (0 or 1)

- Missing Attribute Values: Yes