Data Modelling and Analysis

COMP4030

Coursework 2022 CW2 Brief

Assessment Name	Coursework 2 – Data Analysis Study	**Weight**	75%
Description and Deliverable(s)	This assignment requires you to work in a pair. You will need to analyse a data set using all the data science steps you have learnt to create and compare classification models. You will write your work up as a joint academic paper with a coursework partner, comparing and analysing your results at every stage of the data analysis and modelling pathway (6 to 8 pages including references and diagrams) as stated in this coursework specification. The paper should be submitted in PDF, using the IEEE template for formatting. The code should be submitted as R script.		
Release Date	Tuesday 1st March 2022		
Submission Date	Monday 9th May 2022 by 3pm		
Late Policy **(University of Nottingham default will apply, if blank)**	Work submitted after the deadline will be subject to a penalty of 5 marks (the standard 5% absolute) for each late working day out of the total 100 marks. Late submission deadline is Friday 13 May 2022. Submissions after this date will only be accepted through the extenuating circumstances process.		
Feedback Mechanism and Date	Written feedback in Moodle on the 6th of June 2022		

Instructions

For this coursework assignment you will need be required to work in pairs to analyse a data set (select one from the three provided or find one of your own choice) using all the data science steps you have learnt to create and compare classification models.

You will write your work up as a joint academic paper with your coursework partner, comparing and analysing your results at every stage of the data analysis and modelling pathway.

You will need to present your paper in an IEEE format using a template from here: https://www.ieee.org/conferences/publishing/templates.html

Your paper should be between 6 to 8 pages (including tables, diagrams and references as appropriate) and submitted as a PDF. The diagrams table and diagrams should add value to the writing. Diagrams are preferrable to tables.

Your paper should be organised into 8 parts:

1. Title and Abstract (2.5%)
2. Introduction to the data set and research question(s) (5%)
3. Literature Review – covering a few key methods adopted by other researchers who used this or a similar dataset (5%)
4. Methodology – including a justification for your selected approaches for data analysis and pre-processing and data classification. (10%)

5. Results from each of the stages – data analysis, pre-processing and classification (20%) Please note at each partner in the pair should use a different approach for each stage.
6. Discussion - comparing your results (partners in pair) and also with other results from previous research on the dataset as noted in your literature review (25%)
7. Conclusions and recommendation for future research (10%)
8. References (2.5%)

Code Submission

Please include all your code as an R script which the be run to generate your results (20% = each person in the pair will be marked individually on this) as a separate file in additional to the paper.

The ultimate aim of this coursework is to give you first-hand experience on working with a relatively large and real data set, getting experience of the first stages of data description, exploratory data analysis to the later stages of knowledge extraction and classification/prediction.

Please note that you need to include a contributions section in the paper to clearly specify which person worked on what aspects of the paper.

Datasets

You can choose to work on one of the following datasets:

1. **Wine Data Set**

https://search.r-project.org/CRAN/refmans/HDclassif/html/wine.html

Data Set Information:

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Format: A data frame with 178 observations on the following 14 variables:

Class The class vector, the three different cultivars of wine are represented by the three integers : 1 to 3.
V1 Alcohol
V2 Malic acid
V3 Ash
V4 Alkalinity of ash
V5 Magnesium
V6 Total phenols
V7 Flavanoids
V8 Nonflavanoid phenols
V9 Proanthocyanins
V10 Color intensity
V11 Hue
V12 OD280/OD315 of diluted wines
V13 Proline

2. **Breast Cancer Wisconsin (Diagnostic) Data Set**

https://search.r-project.org/CRAN/refmans/mlbench/html/BreastCancer.html

Data Set Information:

The objective is to identify each of a number of benign or malignant classes. Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself. Each variable except for the first was converted into 11 primitive numerical attributes with values ranging from 0 through 10. There are 16 missing attribute values. See cited below for more details.

Format A data frame with 699 observations on 11 variables, one being a character variable, 9 being ordered or nominal, and 1 target class.

[,1]	Id	Sample code number
[,2]	Cl.thickness	Clump Thickness
[,3]	Cell.size	Uniformity of Cell Size
[,4]	Cell.shape	Uniformity of Cell Shape
[,5]	Marg.adhesion	Marginal Adhesion
[,6]	Epith.c.size	Single Epithelial Cell Size
[,7]	Bare.nuclei	Bare Nuclei
[,8]	Bl.cromatin	Bland Chromatin
[,9]	Normal.nucleoli	Normal Nucleoli
[,10]	Mitoses	Mitoses
[,11]	Class	Class

3. **Pima Indians Diabetes Dataset**

https://search.r-project.org/CRAN/refmans/hhcartr/html/pima.html

Description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. 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.

Format : A data frame with 768 rows and 8 variables and one output:

Pregnancies Number of times pregnant	Number of times pregnant
Glucose Plasma	glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function

Age	Age (years)
Outcome	target variable, whether patient had diabetes, 268 of 768 are 1, the others are 0 (1 = yes; 0 = no)

Assessment Criteria

The main assessment criteria for the paper are:

Section	Weight-ing %	Criteria
Title and Abstract	2.5	Are the title and abstract appropriately reflective of the content of the paper?
Introduction to the data set and research question(s)	5	Have the data set and research question(s) been clearly defined?
Literature Review – covering a few key methods adopted by other researchers who used this or a similar dataset	5	Have relevant papers been discussed and their approaches and results succinctly described?
Methodology – including a justification for your selected approaches for data analysis and pre-processing and data classification. Please note at each partner in the pair should use a different approach for each stage.	10	Have at least two different approaches for each stage been suggested? Have the selected approaches been clearly discussed and justified? Are they appropriate to the problem at hand?
Results from the different approaches applied at each of the stages – data analysis, pre-processing and classification	20	Were the techniques applied correctly? Have the results from at least two alternative approaches been included at each stage? Have suitable diagrammatic representations of the results been included?
Discussion - comparing your results (partners in pair) and also with other results from previous research on the dataset as noted in your literature review	25	Have the findings been interpreted in an appropriate manner? Have the results been compared in a critical manner? Have the results from the different techniques/approaches also been compared to results from other research studies on this dataset?
Conclusions and recommendation for future research	10	Is there is a good summary of the work? Is there consideration of the shortcomings of the work? Are there any suggestions regarding how the techniques could be further combined in new and interesting ways?
References	2.5	Have appropriate references been included and cited correctly?

R code	20	Is the code well commented and easy to follow? Is it consistent (i.e. consistent names for variables, functions, etc.)? does it use informative names for variables and functions? Does it give the results as stated in the paper?

Module study expectations

Activity	Per Week	Total Hours
Lecture – delivery key material	2 × 12	24
Lab sessions	2 × 12	24
Self-study – review lecture content and read associated background materials	6 × 12	72
Coursework (75%)		**60**
Lab submission Preparation (25%)	6.7 x 3	20
Total (20 credits)		**200**

References:

Jain, A., Bhandari, N.S. and Jain, N., 2018, February. Essential elements of writing a research/review paper for conference/journals. In *2018 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS)* (pp. 131-136). IEEE. (Paper on Moodle)

Other resources for this coursework assignment:

Reading list on Moodle

Materials covered and referenced in the lectures and lab sessions.