B.Sc. in IT Management
B.Sc in Computing
Higher Diploma
Enterprise Database Technologies
CA 2 - INDIVIDUAL
This CA is worth 25%
Upload by 20th April 2018

Overall, this CA is designed to assess your skills and abilities in determining the optimal machine learning model for your data-set. Additionally, it will also assess your skills in using the Apriori Algorithm in Association Rules Discovery. There are **two** sections to the report.

Section 1 – Classifiers. Use the file provided on Moodle.(80%)

If your student number ends in an even number **use Data-set 1.**

If your student number ends in an odd number **use Data-set 2.**

- **Data Set 1** Name: Autistic Spectrum Disorder Screening Data for Adults
- **Data Set 2** Name: Autistic Spectrum Disorder Screening Data for Children

The objective of the ML project is to detect patients that have ASD.

Section 2 - A-Priori Algorithm Manual Exercise (20%).

This CA is on an individual basis. A brief viva/presentation of findings may be required as part of this work. Any work not directly your own must be referenced. NB: Institute plagiarism rules apply to this and every CA.

The assignment will be checked through the Turnitin Plagiarism Prevention system, for identifying unoriginal material, copied (without reference to the source) from an electronic source on the Internet, electronic libraries, other assignments.

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

The Weka Data mining tool should be used for this assignment.

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.

The report should be no more than 6 pages (excluding appendices).

Appendices should be use to provide the evidence of the work carried out e.g. model configurations, model outputs, Experimenter outputs.

Section 1

You have been asked to identify the best performing Classification model on your dataset using **Weka**. The objective of the ML project is to detect patients that have ASD. The tasks you should carry out are listed below. You should write up key decisions made and any significant findings.

Load and Analyse the dataset

1. Using Weka, you should review the data before modelling. Assess the summary statistics for each attribute identifying the scale, missing values, unusual values. You should evaluate the distribution of each numeric attribute by visual inspection and the interactions between attributes for correlations. You should write up any significant findings and their implications.

2. **Prepare a number of views (formats) of the dataset** (i.e. save different formats)
   a. The original dataset file ([DatasetName]**Original.arff**).
   b. Normalised view e.g. Min-Max or Z-Score (call the dataset [DatasetName]**normalised.arff**)
   c. Standardised view that rescales the data so that the each predictor variable has a mean of 0 and a standard deviation of 1 (suitable for ML Algorithms like Logistic Regression or Naïve Bayes) (call the dataset [DatasetName]**standardised.arff**)
   d. With regard to attributes with missing data or unusual data, impute the average\mode value for each attribute (call the dataset [DatasetName]**missing.arff)**

   [Use appropriate filters in Weka to accomplish this]

3. **Attribute Selection (feature selection)**
   You should determine the optimal number of predictor variables to use in the modelling stage of the assignment. Use up to three attribute selection methods to determine the attributes. Explain your decisions.

4. Use kNearestNeighbour (IBk) Classifier on the dataset. Try it out for k= 3, 5, 7 and 15. Comment on your findings. Now determine the optimal value of k. Does it perform better compared to the previous models. Explain your reasoning decisions.

5. Identify two suitable Machine Learning (ML) Algorithms (**not listed in 6.**) you would like to use in the evaluation. One must be a suitable Ensemble Method. Write a concise narrative on each ML algorithm chosen outlining how it functions as a classifier.

6. Carry out an initial evaluation on the following ML algorithms:
   - rules.ZeroR --baseline algorithm
   - rules.JRip
   - bayes.NaïveBayes
   - functions.SMO
   - lazy.IBk - using the k value you have determined in **4.**
   - trees.J48
   - + 2 ML algorithms you have identified in **5.**

You should also load the 4 .arff files into Experimenter (e.g. data.arff, normalised.arff, standardised.arff, missing.arff). As part of your write-up comment on

- The performance of the ML algorithms.
- The results on using the different data-sets.
- The best performing algorithm(s).
- Significant findings and supporting evidence.

[It is suggested that you use the **Weka Experimenter** to complete this task.]

7. **Final Version of the Model and Present Results**

   Create a final version of your model using the whole dataset by using test option *Use Training set.* Classify the unseen cases.

Section 2

In this section, you will carry out unsupervised association rules discovery mining in a market basket analysis exercise.

Consider the following transaction database from a local supermarket. Each row represents a single transaction in which the specified items have been purchased. You must solve this problem manually by hand-tracing the Apriori algorithm discussed in the lectures rather than using a data mining tool or program. Excel can be used to assist you.

Trans ID	**Items Purchased**
1	NIKON CAMERA, MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE;
2	MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE, PS4 GTA GAME, PS4 FIFA 19 GAME;
3	NIKON CAMERA, SHOOT TRIPOD, PS4 FIFA 19 GAME, CHARGER, PS4 CONTROLLER;
4	MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE, PS4 GTA GAME, PS4 CONTROLLER;
5	PS4 CONSOLE, PS4 GTA GAME, IPAD, CHARGER, AMAZON ECHO;
6	NIKON CAMERA, MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE, PS4 GTA GAME, AMAZON ECHO;
7	NIKON CAMERA, PS4 CONSOLE, PS4 GTA GAME, IPAD, AMAZON ECHO;
8	MICRO SD CARD, FITBIT, PS4 CONTROLLER, AMAZON ECHO;
9	SHOOT TRIPOD, PS4 CONSOLE, IPAD, AMAZON ECHO;
10	NIKON CAMERA, MICRO SD CARD, PS4 CONSOLE, PS4 GTA GAME, PS4 CONTROLLER;
11	SHOOT TRIPOD, PS4 CONSOLE, CHARGER, FITBIT, PS4 CONTROLLER;
12	MICRO SD CARD, SHOOT TRIPOD, PS4 GTA GAME, PS4 CONTROLLER;
13	MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE, IPAD;
14	NIKON CAMERA, MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE;
15	SHOOT TRIPOD, CHARGER, FITBIT, FITBIT WRIST BANDS;
16	NIKON CAMERA, PS4 GTA GAME, IPAD, CHARGER, AMAZON ECHO;
17	CHARGER, PS4 CONTROLLER, AMAZON ECHO;
18	NIKON CAMERA, MICRO SD CARD, PS4 CONSOLE, CHARGER, PS4 CONTROLLER;
19	PS4 CONSOLE, PS4 GTA GAME, PS4 CONTROLLER;
20	MICRO SD CARD, SHOOT TRIPOD, PS4 CONSOLE, PS4 GTA GAME, CHARGER;

a. Applying the **Apriori** algorithm to this dataset with minimum support of 30% (i.e. $\varphi\ (phi) = 6$ find all the frequent itemsets in the data set. For each step in the algorithm, give the list of frequent itemsets that satisfy minimum support (i.e., for each iteration ***i***, give the set $\boldsymbol{L_i}$ along with the support values for the frequent itemsets).

b. From the frequent itemsets you have identified, generate and identify the top 10 association rules that maximise support and confidence of the rule.

c. Discuss and interpret the significant rules you have identified.

Data Set 1 Name: Autistic Spectrum Disorder Screening Data for Adult

Abstract: Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioural features (AQ-10-Adult) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science.

Task: Classification
Attribute Type: Categorical, continuous and binary
Area: Medical, health and social science
Does your data set contain missing values? Yes
Number of Instances (records in your data set): 704
Number of Attributes (fields within each record): 21
Relevant Information: For Further information about the attributes/feature see below table.
Attribute Information:

Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Autism - Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Country of residence	String	List of countries in text format
used_app_before	Boolean (yes or no)	Whether the user has used a screening app
result	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
age_desc	String	Age description
relation	String	Who is completing the test e.g. Parent, self, caregiver, medical staff, clinician ,etc.
Class/ASD	Boolean (yes or no)	The Target Variable- whether the patient is diagnosed with ASD

Source: Fadi Fayez Thabtah
Department of Digital Technology
Manukau Institute of Technology,
Auckland, New Zealand

Relevant Papers:

1) Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

2) Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

3) Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press)

Data Set 2 Name: Autistic Spectrum Disorder Screening Data for Children

Abstract: Autistic Spectrum Disorder (ASD) is a neurodevelopment condition associated with significant healthcare costs, and early diagnosis can significantly reduce these. Unfortunately, waiting times for an ASD diagnosis are lengthy and procedures are not cost effective. The economic impact of autism and the increase in the number of ASD cases across the world reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, a time-efficient and accessible ASD screening is imminent to help health professionals and inform individuals whether they should pursue formal clinical diagnosis. The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the efficiency, sensitivity, specificity and predictive accuracy of the ASD screening process. Presently, very limited autism datasets associated with clinical or screening are available and most of them are genetic in nature. Hence, we propose a new dataset related to autism screening of adults that contained 20 features to be utilised for further analysis especially in determining influential autistic traits and improving the classification of ASD cases. In this dataset, we record ten behavioural features (AQ-10-Child) plus ten individuals characteristics that have proved to be effective in detecting the ASD cases from controls in behaviour science.

Task: Classification
Attribute Type: Categorical, continuous and binary
Area: Medical, health and social science
Does your data set contain missing values? Yes
Number of Instances (records in your data set): 292
Number of Attributes (fields within each record): 21
Relevant Information: For Further information about the attributes/feature see below table

Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Age	Number	Age in years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Autism - Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Country of residence	String	List of countries in text format
used_app_before	Boolean (yes or no)	Whether the user has used a screening app
result	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
age_desc	String	Age description
relation	String	Who is completing the test e.g. Parent, self, caregiver, medical staff, clinician ,etc.
Class/ASD	Boolean (yes or no)	The Target Variable- whether the patient is diagnosed with ASD

Source: Fadi Fayez Thabtah

Department of Digital Technology
Manukau Institute of Technology,
Auckland, New Zealand

Relevant Papers:

1) Tabtah, F. (2017). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

2) Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th, 2017].

3) Thabtah, F. (2017). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. To Appear in Informatics for Health and Social Care Journal. December, 2017 (in press)