COMP3425 Data Mining S1 2024

Assignment 2

Maximum marks	**100**
Weight	**25% of the total marks for the course**
Length	**Maximum of 10 pages, excluding cover sheet, bibliography and appendices.**
Layout	**A4 margin, at least 11-point type size, use of typeface, margins and headings consistent with a professional style.**
Submission deadline	**9:00am, Monday, 6 May**
Submission mode	**Electronic, via Wattle**
Estimated time	**15 hours**
Penalty for lateness	**100% after the deadline has passed**
First posted:	**25th March, 9:00 AM**
Last modified:	**9th April 2:00 PM**
Questions to:	**Wattle Discussion Forum**

This assignment specification may be updated to reflect clarifications and modifications after it is first issued. -It is strongly suggested that you start working on the assignment right away. You can submit as many times as you like. Only the most recent submission at the due date will be assessed.

In this assignment, you are required to submit a single **report** in the form of a PDF file. You may also attach supporting information (appendices) as one or more identified sections at the end of the same PDF file. Appendices will not be marked but may be treated as supporting information to your report. Please use a **cover sheet** at the front that identifies you as author of the work using your u-number and name and identifies this as your submission for COMP3425 Assignment 2. The cover sheet and appendices do not contribute to the page limit.

You are expected to write in a style appropriate to a professional report. You may refer to http://www.anu.edu.au/students/learningdevelopment/writing-assessment/report-writing for some stylistic advice. You are expected to use the question and sub-question numbering in this assignment to identify the relevant answers in your report.

No particular layout is specified, but you should use no smaller than 11-point typeface and stay within the maximum specified page count. Page margins, heading sizes, paragraph breaks and so forth are not specified but a professional style must be maintained. Text beyond the page limit will be treated as non-existent.

This is a single-person assignment and should be completed **on your own**. **The use of any Generative AI tools is not permitted.** Make certain you carefully reference all the material that you use, although the nature of this assignment suggests few references will be needed. It is unacceptable to cut and paste another author's work and pass it off as your own. Anyone found doing this, from whatever source, will get a mark of zero for the assignment and, in addition, CECC procedures for plagiarism will apply.

No particular referencing style is required. However, you are expected to reference conventionally, conveniently, and consistently. References are not included in the page limit. Due to the context in which this assignment is placed, you may refer to the course notes or course software where appropriate (e.g. "*For this experiment Rattle was used*"), without formal reference to original sources, unless you **copy text or images** which always requires a formal reference to the source. You do not need to reference this specification.

An assessment rubric is provided. The rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work.

Your submission will be treated confidentially. It will be available to ANU staff involved in the course for marking. It may be shared, de-identified, as an exemplar for other students.

Task

You are to complete the following exercises, using **the supplied data set**. For simplicity, the exercises are expressed using the assumption that you are using Rattle, however you are free to use R directly or any other data mining platform you choose that can deliver the required functions. You should describe the methods used in terms of the language of data mining, not in the terms of commands you typed or buttons you selected. You are expected, in your own words, to interpret *selected* tool output in the context of the learning task. Write just what is needed to explain the results you see.

1. Platform

Briefly describe the platform for your experiments in terms of memory, CPU, operating system, and software that you use for the exercises. If your platform is not consistent throughout, you must describe it for each exercise. This is to ensure your results are reproducible.

2. Data

(a) In your own words, briefly describe the purpose and means of data collection.

(b) Look at the pairwise correlation amongst ordinal variables using Pearson product-moment correlation. Qualitatively describe the pairwise correlations amongst *A1* and each of the variables *RA2_a, RA2_b RA2_c, RA2_d and RA2_e.* **Explain** what you see in terms of the **meaning of the data**.

3. Association mining: What factors affect satisfaction with the country's future?

A1 of the survey asks respondents how they feel about the way the country is heading. Your task is to use association mining to find out which factors might be associated with a person's response to *A1*.

(a) Generate association rules, adjusting *min_support* and *min_confidence* parameters as you need. **What parameters** do you use? Bearing in mind we are looking for insight into what factors affect ~~*RA1*~~ A1, **find 3 interesting rules**, and explain both **objectively** and **subjectively** why they are interesting.

(b) Comment on whether, in general, association mining could be a **useful technique** on this data.

4. Study a simple classification task

Restore the dataset to its original distributed form, if you made any changes above.

Aim to build a model to classify how people voted in the referendum, relying on the variable *RB1.*

To do this, define a new variable *Ass2_voted* to map those who voted Yes to TRUE, those who voted No to FALSE and all others to NA. For example, in Excel you can create a new column, give it the header *Ass2_voted*, enter the following formula to row 2, then copy into all other rows:

```
=IF(AS2=1,TRUE, IF(AS2=2, FALSE,"NA"))
```

Now remove or otherwise ignore all rows in which people responded to *RB1* with something other than *yes* or *no*. Use ~~*RB1*~~ *Ass2_voted* as the target class and set every other variable (except *srcid*) as Input (independent). Remove (or Ignore) variable *RB1* itself.

Using sensible defaults for model parameters is fine for this exercise where we aim to compare methods rather than optimise them. You may choose to remove (or Ignore) some other variables from the data. If you do so, explain what they were and why you removed them.

Keep a copy of your modified form of the data for a later purpose.

(a) Why were you asked to Ignore or otherwise remove the original *RB1* variable for this purpose? **Why?** *Hint:* Think how *Ass2_voted* is defined.

(b) Train each of a Linear, Decision tree, SVM and Neural Net classifier, so you have 4 classifiers. *Hint:* Because the dataset is large, begin with a small training set, 20%, and where run-time speeds are acceptable, move up to a 70% training set. **Evaluate** each of these 4 classifiers, using a confusion matrix and interpreting the results in the context of the learning task.

(c) **Inspect** the models themselves where that is possible to assist in your evaluation and to explain the performance results. Which learner(s) performed best and **why**?

5. Predict a Numeric Variable

Return to working with the original form of the data as distributed.

Weight_final_ref is a factor that is calculated to weight each respondent's answers according to the extent to which they represent a demographic portion of the population, based on the values for the variables *p_geography, p_region, .., p_state_sdc* (excluding *weight_final_ref* and *weight_ref*). You are to train a regression tree or a neural net to predict *weight_final_ref*, using those demographic variables as input (excluding *weight_ref).*

(a) Explain which you chose of a regression tree or neural net and **justify** your choice.

(b) Train your chosen model and tune by setting controllable parameters to achieve a reasonable performance. **Explain** what parameters you varied and how, and the values you chose finally.

(c) **Assess** the performance of your best result using the subjective and objective evaluation appropriate for the method you chose and **justify** why you settled with that result.

6. More Complex Classification

Return to using the modified data you created for Question 4.

Here you are going to work harder to attempt to build a great model to classify *Ass2_voted*. *Hint:* Ensure variable RB1 is not used.

(a) **Explain** how you will partition the available dataset to train and validate classification models in (b) to (d) below.

(b) Train a Decision Tree Classifier. You will need to adjust default parameters to obtain optimal performance. **State** what parameters you varied and (briefly) their effect on your results. **Evaluate** your optimal classifier using the **error matrix**, **ROC**, and any quality information specific to the classifier method.

(c) Proceed as for (b) above, training and discussing an SVM Classifier instead of a Decision Tree.

(d) Proceed as for (b) above, training and discussing a Neural Net instead of a Decision Tree.

7. Clustering

Restore the dataset to its original distributed form.

For clustering, use 5 variables of your choice from *RD7_a* to *RD7_j* . Ignore all the other variables.

Rescale the variables to fall in the range [0-1] prior to clustering. Use the full dataset for clustering (i.e. do not partition into train-test sets).

Experiment with clustering using the k-means algorithm by building cluster models for each of *k*= 2, 5, $\sqrt{\frac{n}{2}}$ (the latter is a recommended default for dataset of size *n*) clusters. Choose your preferred *k* and its cluster model for k-means to answer the following.

(a) **Justify** your choice of *k* as your preferred (*Hint:* have look at parts b-d below for each cluster model).

(b) **Calculate** the sum of the within-cluster-sum-of-squares for your chosen model. The *within-cluster-sum-of-squares* is the sum of the squares of the Euclidean distance of each object from its cluster mean. **Discuss** why this is interesting.

(c) Look at the cluster centres for each variable. Using this information, **discuss** qualitatively how each cluster differs from the others.

(d) Use a scatterplot to plot (a sample of) the objects projected on to each combination of 2 variables with objects mapped to each cluster by colour (*Hint:* The Data button on Rattle's Cluster tab can do this). **Describe** what you can see as the major influences on clustering. **Include** the image in your answer.

8. Qualitative Summary of Findings (*Hint: approx 1/2 page*)

Comparatively **evaluate** the techniques you have used and their suitability or not for mining this data. This should be a *qualitative* opinion that draws on what you have found already doing the exercises above. For example, what can you say about training and classification speeds, the size or other aspects of the training data, or the predictive power of the models built? Finally, what else would you **propose** to investigate as a follow-up to your work presented here?

Assessment Rubric COMP3425 Data Mining

This rubric will be used to mark your assignment. You are advised to use it to supplement your understanding of what is expected for the assignment and to direct your effort towards the most rewarding parts of the work. Your assignment will be marked out of 100, and marks will be scaled back to contribute to the defined weighting for assessment of the course.

Review Criteria	Max Mark	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
1. Platform & 2. Data	10	9-10 1.Platform description complete (memory, CPU, operating system, software). 2a Demonstrates understanding of the purposes and process sufficient to frame report. 2b All correlations for mentioned variables clearly explained in terms of the data semantics, in the correct directions and for correct or plausible domain reasons.		7-8 1. Platform description complete (memory, CPU, operating system, software). 2a Clear description of the the data domajn. 2b Partially clear and correct explanation in terms of data semantics	5-6 1. Platform description complete (memory, CPU, operating system, software). 2a Attempt but unclear 2b Partial description of variables or unclear 2b Partial explanation in data context	0-4 1. Platform description incomplete. 2a Incomplete or faulty 2b Description unrelated to correlation of variables. 2b Explanation unrelated to data source

Criteria	Max	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
3. Association mining	10	9-10 a. Answers demonstrate deep understanding of association mining, by the careful selection of interesting and differentiated rules and clear rationale for interestingness. b. Comment shows original and insightful analysis of association mining on the problem.		7-8 a Support and confidence clear a 3 rules given a objective interestingness is given for all 3 a subjective interestingness attempted b Comment makes sense.	5-6 a Support or confidence not clear a < 3 rules given a objective interestingness is incomplete a subjective interestingness is incomplete b Comment cursory.	0-4 Required information not provided and/or incorrect or misleading, demonstrating lack of engagement with the problem
4. Simple classification	10	9-10 Explanation of *Ass2_voted* demonstrates generalised understanding of issue. Deep understanding of the 4 models demonstrated thorough analysis of performance on the task.		7-8 a correctly explains why *Ass2_voted* required removal of RB1. b 4 confusion matrixes given b confusion matrixes explained in terms of the data and the method and the model learnt. c evidence of understanding what the models are doing c reasoning for comparative performance demonstrating understanding of the methods behind them	5-6 a partially explains why definition of *Ass2_voted* required removal of *RB1* b 4 confusion matrixes given b confusion matrixes explained at face value only c partial understanding of learnt models c comparative performance only cursorily presented c reason for comparative performance is shallow	0-4 a inadequate explanation b confusion matrix missing or misunderstood. c Interpretation of confusion matrix missing or faulty c little understanding of what the models are doing c missing or unexplained comparative analysis

Criteria	Max	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
5. Prediction	20	17-20 Approach to problem demonstrates serious effort to produce good results and a deep understanding of the relative benefits of the 2 methods in the context of the problem domain. Results are interpreted in the context of the problem domain.	14-16 a justification for choice shows understanding of the comparative benefits of each and extensive experiments. b parameter variations shows a combination of experimentation and understanding of the parameters c several subjective and objective evaluation measures used as appropriate to method including synthesised evaluation c justification for stopping demonstrates awareness of appropriateness of best result and scope of potential for further improvement	12-13 a justification for choice shows understanding of the comparative benefits of each and experiments with performance. b parameter variations shows a combination of experimentation and understanding of the parameters. c multiple subjective and objective evaluation measures used as appropriate to method c justification for stopping demonstrates awareness of appropriateness of best result	10-11 a justification for choice shows some understanding of the comparative benefits of each or experiments with performance. b parameter variation demonstrates some experimentation c cursory evaluation given c justification for stopping perfunctory	0-9 a weak justification for choice b variation insufficient c evaluation fails to demonstrate effort or understanding of evaluation c justification for stopping effectively absent

Criteria	Max	Exemplary	Excellent	Good	Acceptable	Unsatisfactory
6. Complex Classification	30	26-30 Exemplary use of classification methods with comprehensive and fit-for-purpose performance analysis on the problem that includes meaningful reflection over the three methods.	22-25 a explanation sound b,c,d parameter variation clear and extensive demonstrating understanding of effect in all 3 methods b.c.d error matrix and ROC correctly interpreted in all 3 methods b,c,d extensive use of specific evaluation methods and significance clearly explained in all 3 methods	18-21 a explanation sound b parameter variation clear and sufficient for good results b error matrix correctly interpreted b ROC correctly interpreted b some specific evaluation methods used c parameter variation clear and sufficient for good results c error matrix correctly interpreted c ROC correctly interpreted c some specific evaluation methods used d parameter variation clear and sufficient for good results d error matrix correctly interpreted d ROC correctly interpreted d some specific evaluation methods used	15-17 a satisfactory approach to dataset partitioning b parameter variation perfunctory b error matrix given b ROC given b few specific evaluation methods used well c parameter variation perfunctory c error matrix given c ROC given c few specific evaluation methods used well d parameter variation perfunctory d error matrix given d ROC given d few specific evaluation methods used well	0-14 a explanation incorrect or unsound use of training/testing/validation data b no parameter variation b no error matrix b no or faulty ROC b specific evaluation methods missing c no parameter variation c no error matrix c no or faulty ROC c specific evaluation methods missing d no parameter variation d no error matrix d no or faulty ROC d specific evaluation methods missing
Criteria	Max	Exemplary	Excellent	Good	Acceptable	Unsatisfactory

7. Clustering	10	9-10 The application of k-means algorithm to the dataset and its evaluation demonstrates exemplary understanding of the algorithm, its evaluation, and its limitations. Suitable evaluation methods or clustering experiments in addition to those required here may be used.		7-8 a Convincing justification for k b Measure calculated correctly. Discussion recognises value and limitations c Discussion on centres reflects numeric results and emphasises the interesting parts that relate to the significance in domain terms d Correct scatterplot included and description shows understanding linked to data domain	5-6 a Justification offered but not clear or unconvincing b Measure calculated correctly c Discussion on centres reflects numeric results d Correct scatterplot included. Attempt at influences.	0-4 Clustering experimentation and discussion inadequate
8. Qualitative Summary	10	9-10 Many aspects of evaluation are discussed and a clear conclusion is drawn, with direct reference to the purpose of the data collection. Proposal for further investigation demonstrates creativity and thoughtful engagement with the problem, clearly building on the work reported.	8 A clear conclusion is drawn from the work reported and a defended proposal for further investigation is proposed, with clear links to both the work reported and the domain of application.	7 A rounded, balanced summary of the work is presented with a justified proposal given.	5-6 A summary of the work is presented and a proposal made.	0-4 Answer does not demonstrate adequate engagement with the problem nor a qualitative understanding of the work reported.