STAT462 – Data Mining

Assessment 11 – Report 3

Attribute	Details
Individual/Group	Group
Course Learning Outcomes (Relevant to this assessment)	The course Learning Outcomes demonstrated by successful completion of the task below include: • Identify suitable algorithms to address specific research questions in tree-based classification and cluster analysis. • Implement, adapt and apply algorithms in R to solve research questions, and accurately interpret their results. • Critically compare and reflect on the limitations and performance of the algorithms under consideration. • Compose comprehensive scientific and technical reports using R Markdown to communicate your analysis and findings effectively.
Conditions	• Your report must be a single .pdf document created from R Markdown. • Once the due date and time has passed no resubmissions are allowed, without permission from the facilitator.
Submission	See LEARN for specific requirements and deadline.
Length	No specific length is expected but there is a strict maximum of 20 pages in pdf form. The report is allowed to be much shorter than this without penalty.
Weighting	15%
Maximum grade	100

Assessment task

The objective of this assessment is to investigate a dataset and predict a categorical response (classification). You will load datasets, do some cleanup and feature engineering, train and run classification algorithms, and interpret your findings in a report. You will be graded on all parts of this procedure.

Practical application

This is the same as for the previous two reports.

Detailed instructions

In this assignment you will analyse a dataset containing the chemical analysis of wine samples, and we will do both tree-based classification and cluster analysis on this dataset.

Note: Usually we need to be very careful not to mix up cluster analysis (unsupervised) and classification (supervised), but in this exercise we will do both:

- In Part 1 you will train a tree-based classification algorithm to detect the "cultivar" (the breed of wine used) of a given wine sample (there are three distinct cultivars in this dataset).
- In Part 2 we "pretend" that we do not know these classes, and perform cluster analysis to find out whether there is reason to believe that the data contains more than one cultivar. Of course this is not a meaningful procedure in practice, but this allows us to test out our cluster analysis algorithms and verify them using the (actually known) class labels.

Both questions will use the dataset `wine.csv` which has been split for your convenience into a training and test set in `wine_train.csv` and `wine_test.csv`. In Question A we will use the class label `Class` (the cultivar), in Question B we will disregard it for cluster analysis purposes, and only use it for validation.

A. Classifying wine samples (classification trees)

In this question we will train tree-based classification algorithms to classify samples according to the variable `Class`, which states the cultivar used for the wine.

1. Train a single tree-based classifier on the training set. Use cross-validation to prune this tree suitably. Visualise the classification tree.
2. Prune your tree enough so that you only need two features to make predictions. Visualise your data in these two dimensions, and illustrate the decision tree of your classifier graphically.
3. Fit a bagged classification tree model and/or a random forest to see whether you can improve on your single tree's performance.

B. Clustering the wine dataset (Hierarchical clustering and k-means)

1. Perform hierarchical clustering on the wine dataset, but do not include the `Class` feature. Group the data into three clusters and check/visualise whether this is a good reconstruction of the (actual) classes recorded in the `Class` feature

2. Do the same using the k-means algorithm, for k=3. For visualisation, you can pick the two features that were sufficient for classification in question A, and plot datapoints in these two dimensions, comparing actual classes and predicted cluster labels.

3. You will likely not get great results, because your features vary on very different orders of magnitude (for example, `Nonflavanoid.phenols` is mostly between 0 and 1, but `Proline` is in the 1000 range). Normalise all numerical features using either z-score transformation or min-max normalisation, which will bring them into comparable orders of magnitude. Then repeat parts 1. and 2., and see whether your results improve.

Recommended structure

All answers need to be referenced to the original question, and we recommend you go through the questions in the order in which they are presented above.

Research and referencing

Additional research and referencing is not required to complete this assessment, but any ideas that are not your own need to be cited using APA 7th edition.

Presentation guidelines

- Your report must be a single .pdf document created from R Markdown.
- Your report needs to contain and display all R code that needs to be run to get your results.
- All code needs to be sufficiently commented
- Full report (including all figures and code) must not exceed 20 pages in .pdf, but it can be much shorter (the instructor's model solution is about 12 pages long, and there will be no penalty for shorter reports with the same content).

How your report will be assessed

This rubric provides a general guide to how your work will be evaluated. Not all criteria apply to every question—some may be more relevant than others. Focus on producing clear, well-documented, and accurate work.

Grading Overview

Grade range 50%-65%

- You have attempted the assignment, but in a limited way that might not fully demonstrate understanding of the concepts involved.

Grade range 65%-79%

- You have completed most of what was required and did nearly everything correctly.

Grade range 80%-100%

- You have done all that was required and everything is correct.

General Guidelines

- **Code quality and documentation**: Your code should be readable, well-commented, and functional. Errors will result in reduced marks.
- **Statistical accuracy**: Ensure calculations (e.g., regression coefficients, R^2, confidence intervals) are correct and appropriately rounded (in the context of this course, accuracy of 3 decimal places is sufficient unless stated otherwise).
- **Interpretation of results**: Provide logical conclusions supported by data.
- **Data visualisation**: Figures should be clear, well labelled, and appropriately formatted.

- **Report quality and presentation**: The report should be well-structured and professional, with proper formatting, grammar, and clear explanations, and not longer than necessary.
- **Model comparison (if applicable)**: If required, discuss performance and insights. If a specific model is assigned, implement and explain it clearly.

Detailed grading expectations

Skills	Expectations for 50-65%	(Additional) Expectations for 65-79%	(Additional) Expectations for 80-100%	Approximate Weighting
Communicating clearly via text	Text is in understandable English. For most questions, a few short sentences will be adequate unless stated otherwise.	Uses adequate technical jargon and concise language. "Generic garbage" as sometimes produced by GenAI will lead to detractions here.	Writing is clear, concise, and well-structured. Uses precise terminology and presents complex ideas in an accessible way, demonstrating strong communication skills.	Necessary prerequisite. Work cannot be marked if not communicated clearly.
R Markdown and format of submission	Submission is a single file in either .html or .pdf format, generated by R Markdown.	–	–	Necessary prerequisite.
R programming skills	Some attempt to implement or use a statistical algorithm is made.	Code does what it is intended to do. Code is readable, uses comments, and applies suitable naming conventions. Syntax is correct, and code runs without errors.	Code is reasonably efficient and modular, demonstrating best practices in programming. dplyr, ggplot, and statistical learning modules are used appropriately.	**30%**
Knowing and applying statistical learning algorithms	Basic statistical learning task is addressed.	Algorithm results are (when required) put into context, correctly explained, and relevant implications identified.	Demonstrates deep understanding of model limitations and improvements.	**35%**
Interpreting and evaluating algorithm results and output	Some interpretation of the findings is provided, even if not correct.	Correct choice (within constraints of admissible modules) and application of algorithm.	Provides critical analysis, discussing model strengths and weaknesses. Compares results with alternative approaches, considers broader implications, and supports claims with strong reasoning.	**25%**
Communicating clearly via figures	An attempt at visualisation (if required) is made.	Axes are labelled, figure is suitably scaled, and the intention is clearly communicated. Interpretation of the figure is given in either an expressive caption or in the accompanying text.	Choice of visualisation type and colour scheme are highly effective in conveying information.	**10%**

Final Notes

- Not all criteria apply to all questions—focus on what's relevant to each task.
- Marks will be assigned based on correctness, clarity, and depth relative to the question's requirements.
- Follow the instructions carefully and justify key decisions where relevant.

This rubric is a guide to help you understand expectations. If you're unsure about something, refer to the assignment instructions or ask for clarification.