Department of Computer Science

Summative Coursework Set Front Page

Module Title: Data Science Algorithms and Tools

Module Code: CS3DS19

Lecturer responsible: Dr. Carmen Lam

Type of Assignment: Major Coursework

Individual / Group Assignment: Individual

Weighting of the Assignment: 50%

Page limit/Word count: a report of max 4 pages excluding diagrams and graphs
(Times New Roman, 12pt., 1.15 line spacing)

Expected hours spent for this assignment: 10 hours

Items to be submitted on-line through Blackboard Learn:

1. report.pdf,
2. workflow_group.knar (exported KNIME workflow group containing two workflows, one for each task as shown in the figure below).

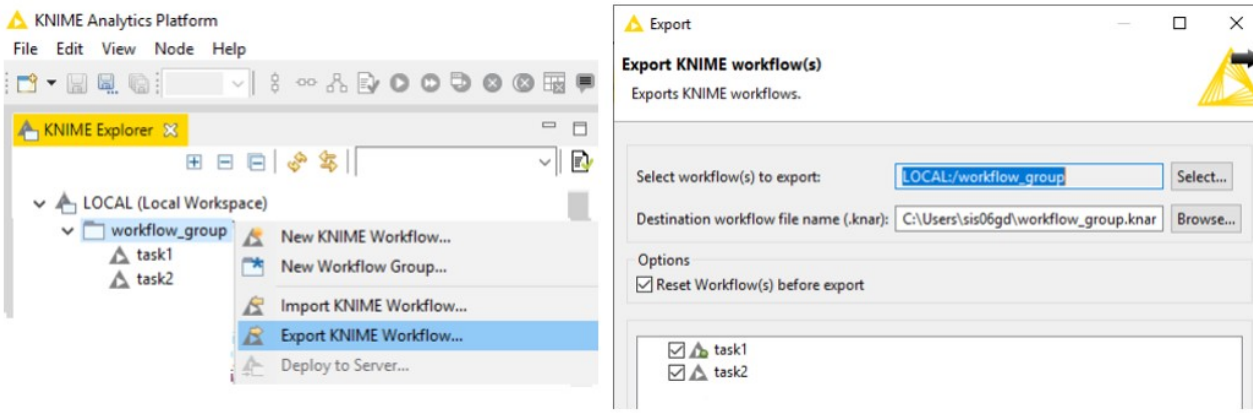

Figure 1: How to export a KNIME workflow group

Work to be submitted on-line via Blackboard Learn by: Friday 25 March 2022 12:00 noon

Work will be marked and returned by: 15 working days after the above deadline

NOTES:

By submitting this work, you are certifying that it is all your sentences, figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work except where explicitly the works of others have been acknowledged, quoted, and referenced. You understand that failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalized accordingly. The University's Statement of Academic Misconduct is available on the University web pages.

If your work is submitted after the deadline, 10% of the maximum possible mark will be deducted for each working day (or part of) it is late. A mark of zero will be awarded if your work is submitted more than 5 working days late. You are strongly recommended to hand work in by the deadline as a late submission on one piece of work can impact on other work.

If you believe that you have a valid reason for failing to meet a deadline then you should complete an Extenuating Circumstances form and submit it to the Student Support Centre before the deadline, or as soon as is practicable afterwards, explaining why.

Assessment classifications

The table below shows what is typically expected of the work to obtain a given mark.

Classification Range	Typically the work should meet these requirements
First Class (>= 70%)	Outstanding/excellent work with correct results, a good presentation of the workflows, code and results, and a critical analysis of the results. An outstanding work will present fully automated solutions based on advanced techniques.
Upper Second (60-69%)	Very good work with partial correct results: most work has been carried out correctly. Some tasks have not been carried out or are not completely correct. The presentation is good, well structured, clear and complete with respect to the work done.
Lower Second (50-59%)	Good work which is missing some significant part of the assignment, and/or with partially correct results. Some tasks have not been carried out. The presentation is, in general, accurate and complete, but it lacks clarity (presentation quality).
Third (40-49%)	Acceptable solutions to limited part of the assignment. Some tasks have not been carried out. Some results may not be complete or technically sound. The presentation is not accurate, complete and lacks clarity.
Pass (35-39%)	Partial solutions to limited part of the assignment. Some tasks have not been carried out. Some results may not be complete or technically sound. The presentation is not accurate, complete and lacks clarity.
Fail (0-34%)	Incomplete solutions to limited part of the assignment. Most tasks have not been carried out with sufficient accuracy. Results may not be correct or technically sound. The presentation is not accurate, complete and lacks clarity.

ASSIGNMENT DESCRIPTION
Major Coursework (50% of module assessment)

This assignment should be carried out using the Data Mining and Machine Learning platform KNIME.
The data file **wine.csv** is required to carry out this assignment and is available in Blackboard.

Task #1 and Task #2 – Data Exploration and Clustering

You are required to perform a clustering analysis for the multidimensional 'wine' data set. The dataset (wine.csv) is obtained from 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 chemical constituents found in each wine. Each data record contains the cultivar ID (1, 2 or 3) and 13 numerical attributes.

Task #1: Clustering without normalization

1.1 Apply Principal Component Analysis (PCA) to generate two-dimensional coordinates and a 2D plot (**plot1**) of the records. The data points in plot1 should be represented with a colour associated with their class label.

1.2 Apply a clustering algorithm to the original data set to generate three partitions. Generate a 2D plot (**plot2**) based on the same PCA projection, similarly to the previous one, where the colour is associated with the cluster ID (use different colours w.r.t. plot1).

1.3 In the report, compare, discuss and explain the differences between plot1 and plot2.

1.4 For the records associated with each cluster, generate a 2D plot (i.e. **plot3a** shows records of cluster_0, **plot3b** shows records of cluster_1, **plot3c** shows records of cluster_2) with colour associated with the class label (same colours of plot1). Visually verify the distribution of class labels in each cluster, report and discuss your observations.

1.5 Select, describe and apply at least one cluster validity measure, report and discuss the results in the report.

The submission for Task #1 must contain two components:

- a report section dedicated to your solution for Task #1 which includes descriptions of the clustering algorithm and the cluster validity measure adopted, descriptions and explanations of the workflows and the node configurations, results and discussions of results.
- any Task #1 KNIME workflow within the group archive.

Task #2: Clustering with normalization

Apply a normalization pre-processing to the data set and repeat the steps of Task #1. Compare the results (plots and cluster validity measure) of Task #1 and Task #2 and discuss how normalization pre-processing affects the results.

The submission for Task #2 must contain two components:

- a report section dedicated to your solution for Task #2 which includes description, workflow and the node configurations of the normalization pre-processing (you do not need to repeat the description of the rest of the workflow that are the same as Task #1), results of Task #2, comparison of Task #1 and Task #2 results (plots and cluster validity measure) and discussions.
- any Task #2 KNIME workflow within the group archive.

CS3DS19 Major Coursework
Marking scheme

		Range for marking	mark
1.	Task #1.1: workflow and node configurations, plot1	0 – 10	
2.	Task #1.2: description of the Clustering algorithm adopted, workflow and node configurations, plot2	0 – 10	
3.	Task #1.3: comparison between plot1 and plot2	0 – 10	
4.	Task #1.4: workflow and node configurations, plot3a, plot3b, plot3c, observations of class labels in each cluster and discussions	0 – 10	
5.	Task #1.5: description of the cluster validity measure adopted, workflow and node configurations, results and discussions	0 – 10	
6.	Task #2: description of the normalization pre-processing, workflow and node configurations	0 – 10	
7.	Task #2: new plots of plot1, plot2, plot3a, plot3b, plot3c	0 – 10	
8.	Task #2: comparison of Task #1 and Task #2 results (plots and cluster validity measure) and discussions	0 – 10	
9.	Quality of the report: overall quality of the document (readability, completeness, presentation quality, references, etc.)	0 – 20	

Total:	**0 - 100**	