# Department of Computer Science

# Summative Coursework Set Front Page

**Module Title:** Data Analytics and Mining

**Module Code:** CSMDM16 / CSMDM21

**Lecturer responsible:** Dr. Carmen Lam

**Type of Assignment:** major coursework

**Individual / Group Assignment:** Individual

**Weighting of the Assignment:** 100%

**Page limit/Word count:** a report of max 10 pages (Times New Roman, 12pt., 1.15 line spacing)

**Expected hours spent for this assignment:** 35 hours

**Items to be submitted on-line through Blackboard Learn:** a single zip archive containing

1. the report (PDF),
2. KNIME workflow images, and
3. KNIME workflow knar archive.

**Work to be submitted on-line via Blackboard Learn by:** **3 December 2021 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.

## MARKING CRITERIA

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 |
|---|---|
| **Distinction (≥=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.<br>- All parts of the assignment are completed correctly,<br>- deep & insightful analysis of the data,<br>- helpful & precise comments,<br>- excellent & compelling presentation of the work. |
| **Merit (60-69%)** | Very good work with mostly 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. |
| **Good (50-59%)** | Good work with solutions to limited part of the assignment. Some significant part of the assignment is missing and/or has partially correct results. The presentation is, in general, accurate and complete, though it may lack some clarity and quality. |
| **Work below threshold standard (40-49%)** | Achievement below the minimum requirements. Solutions limited to some parts 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. |
| **Unsatisfactory standard (<40%)** | 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 (100% of module assessment)

The tasks should be carried out using KNIME, the data analytics and mining tool presented in the lectures. The report should be clearly structured with a separate section for each subtask. Figures and sections must be numbered. References should follow a suitable academic format (https://www.reading.ac.uk/library/finding-info/guides/lib-citing-references.aspx).

In your coursework report, you should include:

1. a brief description of the adopted solutions (data workflows, KDD processes),
2. a brief description of the adopted data mining algorithms,
3. images of KNIME workflows and their presentation (incl. relevant node configurations),
4. the results and their critical analysis.

## Problem #1 – Hierarchical Clustering

You are required to extract and provide some statistics (**task #1**) about the dataset "teeth.csv" (available on Blackboard): these should include the number of records, the number of attributes, the range and mean value of each attribute, and the histogram for each attribute.

You are required to apply hierarchical clustering to all the records in the dataset and to visualize the resulting dendrogram (**task #2**). The dendrogram should report the class labels of the data records at the leaf nodes. The corresponding KNIME workflow should be reported and discussed.

## Problem #2 - Classification

You are required to build and test a classification model for each of the following four datasets:

- the iris dataset: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
- the wine dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
- the breast-cancer-wisconsin dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data
- a dataset of your choice (excluding iris, wine, breast-cancer-wisconsin, teeth)

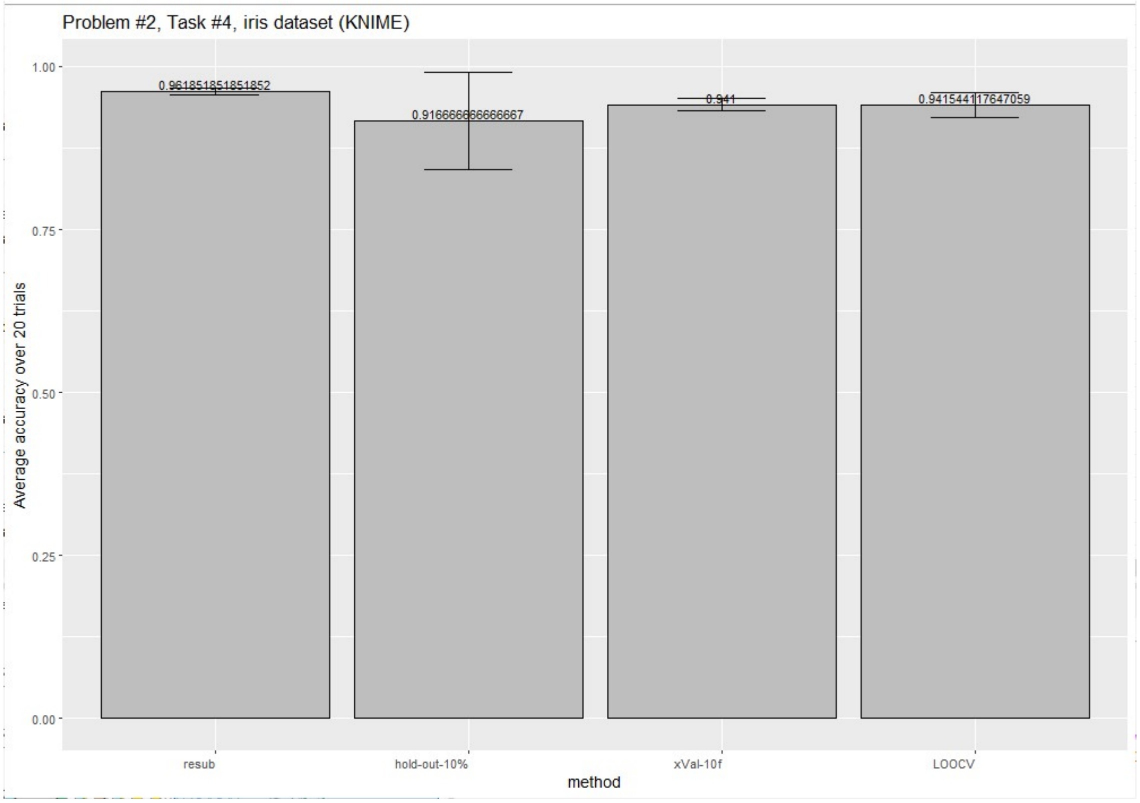
The followings are the descriptions of the corresponding datasets:

- http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names
- https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names
- https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names

You can use any classification algorithm. You are required to carry out a comparison of the accuracy (average and standard deviation over 20 trials) of the four evaluation methods listed below for each of the four datasets using the same classification algorithm. All predictive models must be trained with the same amount of data.

- **resub**: resubstitution error method
- **hold-out-10%**: hold-out method with 10% - 90% partition split
- **xVal-10f**: 10-fold cross-validation method
- **LOOCV**: leave-one-out cross-validation method

The KNIME workflow and the adopted data mining algorithm (**task #3**) must be described and commented in the coursework report just for one of the datasets. The results should be presented and commented (**task #4**), e.g. by means of a bar chart of the average accuracy with standard deviation showing as error bar for each dataset to compare the four methods (an example is shown below).



Finally, overall conclusions of the four methods (e.g. accuracy, performances, how to select between the four methods, etc.) should be provided.

# CSMDM21 Major Coursework
# Marking scheme

| | | available marks | mark |
|---|---|---|---|
| 1. | Problem #1, task #1: statistics on dataset | 0-10 | |
| 2. | Problem #1, task #2: Hierarchical clustering in KNIME (workflow and dendrogram) | 0-10 | |
| 3. | Problem #2, task #3: description of adopted solutions (KNIME workflow) and of the algorithm | 0-10 | |
| 4. | Problem #2, task #4a: compare four methods (dataset 1) | 0-10 | |
| 5. | Problem #2, task #4b: compare four methods (dataset 2) | 0-10 | |
| 6. | Problem #2, task #4c: compare four methods (dataset 3) | 0-10 | |
| 7. | Problem #2, task #4d: compare four methods (dataset 4) | 0-10 | |
| 8. | Problem #2: overall comments and conclusions | 0-10 | |
| 9. | Coursework report (**max 10 pages**) **-** overall presentation quality: structure, readability, completeness, correctness, quality of figures and tables, references, etc. | 0-20 | |

| | | |
|---|---|---|
| **Total** | **0-100** | |