Department of Computer Science
Assessed Coursework Assignment Brief

Module code: CS3DS19

Lecturer responsible: Prof. Giuseppe Di Fatta

Coursework description: Major Coursework

Work to be submitted on-line via Blackboard by 12 noon on: Friday 23 March 2021

Work will be marked and feedback returned by: Prof. Giuseppe Di Fatta

Work will be marked and feedback returned by: within 15 working days of the deadline

This coursework should be submitted on-line through Blackboard Learn.

NOTES:

By submitting this work you are certifying that it is all your own work and that use of material from other sources has been properly and fully acknowledged in the text. You are also confirming that you have read and understood the University's Statement of Academic Misconduct, available on the University web-pages.

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 Information 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
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.
Below Honours Threshold (0-39)	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.

ASSIGNMENT DETAILS: Project for Major Coursework (50%)

The project should be carried out using the Data Mining and Machine Learning platform KNIME.

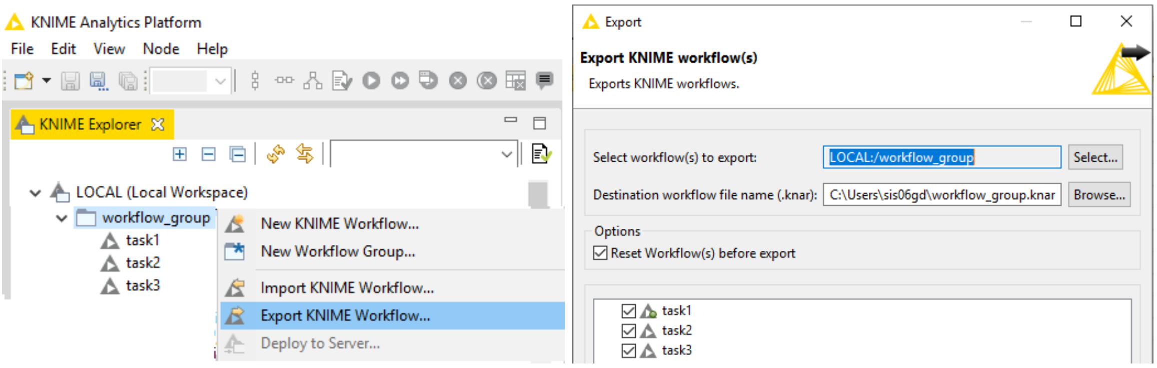
The following data files are required to carry out this assignment and are available in Blackboard:

- wine.csv (data file for tasks 1 and 2)
- training100Ku.csv (data file for tasks 3)
- test1K.csv (data file for tasks 3)

Submission of student work

A report (PDF file), KNIME workflows as a single archive (.knar) and a single file with the prediction results in the required format must be submitted to Blackboard as a **single archive** (.zip) containing a single folder with the following content:

- report.pdf (including sections for the three tasks)
- workflow_group.knar (*) (exported KNIME workflow group containing three workflows, one for each task as shown in the figure below)
- Task3-predictions.csv



Select workflow(s) to export: LOCAL:/workflow_group

Destination workflow file name (.knar): C:\Users\sis06gd\workflow_group.knar

☑ Reset Workflow(s) before export

☑ task1
☑ task2
☑ task3

Figure 1: How to export a KNIME workflow group

Guides on Academic Writing

Guides on academic writing and examples of appropriate and non-appropriate references are available in Bb under “Assessment”.

*(*) **Important: do not include data when you export the KNIME workflow group**. This can be done by selecting “Reset Workflow(s) before you export”*.

Task #1 – 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.

This task has to be carried out two times: with and without normalisation.

Task1.1: Clustering without normalisation

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 to their class label. Apply a clustering algorithm to the 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 to the cluster ID (use different colours w.r.t. plot1), and compare it with plot1. For the records associated to each cluster generate a 2D plot (**plot3a, plot3b, plot3c**) with colour associated to the class label (same colours of plot1): visually verify the distribution of class labels in each cluster.

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

Task1.2: Clustering with normalisation

Apply a normalisation pre-processing to the data set and repeat the steps of the part 1. Compare the new plots and the cluster validity measure with the previous ones.

The submission for Task #1 must contain two components:

- a report section dedicated to your solution for Task #1,
- any T1 KNIME workflow (*) within the group archive.

Task #2 – Comparison of Classification Models

You are required to learn and test classification models for the same ‘wine’ data set used in Task 1. For this Task 2 you need to carry out a performance comparison of TWO different classification algorithms. You should use a 10-fold cross-validation method to estimate the generalisation error. In the report you should briefly describe your selected algorithms, the method used to compare the two algorithms and the results.

The submission for Task #2 must contain two components:

- a report section dedicated to your solution for Task #2,
- any T2 KNIME workflow (*) within the group archive.

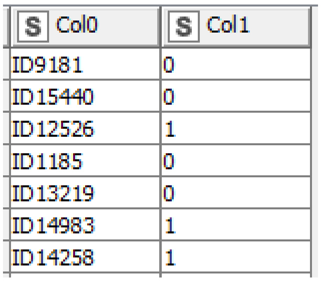
Task #3 – A Data Science Challenge
The Search for God Particle: a Binary Classification Challenge

The CERN's Large Hadron Collider (LHC) typically produces approximately 10^{11} collisions per hour and about 300 (0.0000003%) of these collisions result in a Higgs boson, the so-called God particle. Detecting when interesting particles are produced is an important challenge, which is typically studied by the use of simulations. The data set for this task is related to simulations of collision events, which can be used to train a classification model to distinguish between collisions producing particles of interest (class "signal") and those producing other particles (class "background").

Two data files are provided: the training set (training100Ku.csv) and the test set (test1K.csv). The training set file has 100,000 records, each containing, in this order, 21 numerical low-level attributes, 7 high-level attributes and the class label (signal/background). The low-level attributes are kinematic properties measured by the particle detectors in the accelerator during the experiment. The high-level attributes are computed after the experiment by some complex model (not available) based on the low-level attributes.

The test set has 1,000 records with balanced classes: each record containing a unique record identifier and 21 numerical low-level attributes (the same measurements in the same order as in the training set). The 7 high-level attributes and the class label are not present in the test set. The class label of the test set must be represented by a binary value: 1 corresponds to the class "signal" and 0 to the class "background".

Your goal is to predict the class label for the records of the test set. **The resulting predictions must be submitted as a single CSV file ("Task3-predictions.csv") with only two columns: the record ID and the predicted class, i.e. either 0 (background) or 1 (signal). Please notice that the CSV file should not include any column header.**



S Col0	S Col1
ID9181	0
ID15440	0
ID12526	1
ID1185	0
ID13219	0
ID14983	1
ID14258	1

Figure 2: An example of a part of the file "Task3-predictions.csv". (Please be aware that the data in this example may not correspond to correct predictions.)

You must also include a section in the report to describe the method used to generate the submitted predictions and an estimation of these performance indices: the overall accuracy plus precision, recall and F-measure for the signal class.

In summary, the submission for Task #3 must contain three components:

- a section in the report dedicated to your solution for Task #3,
- any T3 KNIME workflow (*) within the group archive and
- the file "Task3-predictions.csv".

CS3DS19 – Data Science Algorithms and Tools
Major Coursework - Assessment and Feedback Form

		comments and feedback	range for marking	Lecturer's evaluation
1.	Completeness of the submission and quality of the report: overall quality of the document (readability, completeness, presentation quality, etc.)		0-20	
2.	Task #1: description of the Clustering algorithm and the cluster validity measure		0-5	
3.	Task #1: description of the data workflow		0-5	
4.	Task #1: results (10 charts and measures)		0-10	
5.	Task #1: Conclusions and References		0-5	
6.	Task #2: description of the two Classification algorithms adopted		0-10	
7.	Task #2: description of the 10-fold cross-validation method		0-5	
8.	Task #2: experimental results (comparative performance analysis)		0-10	
9.	Task #2: Conclusions and References		0-5	
10.	Task #3: description of the data mining algorithm, the solution (data workflow) adopted and the predicted performance indices.		0-10	
11.	Task #3: prediction results (overall accuracy and three indices for the class "signal": precision, recall and F-measure). These indices will be computed by the lecturer using the submitted file "Task3-predictions.csv".		0-10	
12.	Task #3: Conclusions and References		0-5	

Total	**0-100**	