CSE5DMI 2022 Assignment Two [20 marks]
Assignment Due: 16th Oct 2022 11:59 PM, Sunday (Week 12)

GENERAL DESCRIPTION

This **INDIVIDUAL** assignment consists of **THREE PARTS** and is worth **20%** of the assessment of this subject.

PART I (10 MARKS)

In this part, we are going to build a neural network classifier (NN) for the given dataset. You will be working with a subject/unit dataset taken from a major American university. This University is selective and therefore attracts a student body with relatively high entrance qualifications. The dataset has been partially cleaned but still contains blank or null values. You are to build a NN predictive stream that predicts "At_Risk" (attribute) status, which is a logistic/binary/flag classification target. A detailed list of feature descriptions can be found below.

Column	**Feature Name**	**Feature Description**
A	GRD_PTS_Per_Unit	Grade Points Per Unit
B	**At_Risk (Class label)**	**At-Risk of Failure (Classification Target Variable)**
C	Catalog_NBR	Catalog Number - *Appears to indicate multiple offerings of the same subject, possibly a summer offering and a typical semester offering.
D	GPAO	Grade Point Average in Other Units/Classes *This is the student's overall GPA, in all other subjects, excluding the one captured in this spreadsheet.
E	ANON_INSTR_ID	Anonymous Instructor ID – Indicates the academic or tutor who teaches the student and is likely to mark their work.
F	TERM	Teaching term in which the subject was offered/taught
G	HSGPA	High School Grade Point Average – The student's GPA from high school / secondary education.
H	LAST_ACT_ENGL_SCORE	Last ACT English Score
I	ACT MATH	ACT Mathematics Score
J	ACT READ	ACT Reading Score
K	ACT SCIRE	ACT Science Reasoning Score

L	ACT COMP	ACT Comprehensive Score
M	SEX	Student sex/gender

1. Is the original data ready to be used? [**4 marks**]

   i. If not, state the reason and write a Python script to perform any necessary pre-processing so that the data becomes suitable to be used.

   ii. Briefly describe the pre-processing you carried out with a brief comment in your python script and word document

   iii. Submit the pre-processed data in CSV format

2. Create a NN classifier **[6 marks]**

Requirements

- Split your pre-processed data into training and testing sets in a ratio of 75% to 25%, respectively.
- Report your best classification model and other model results obtained during fine-tuning the following parameters:
  - Number of hidden layers
  - Number of neurons
  - Maximum number of iterations
  - Learning rate
  - Optimizer
- Use Cost Matrix as the primary evaluation measure to represent the model performance.
  - It is worse to classify a student as not at risk when they are at risk (10), than it is to classify a student as at risk when they are not (1). Thus, the cost matrix is given as:

	Predicted not at risks	Predicted at risk
Actual not at risk	0	1
Actual at risk	10	0

- Provide other evaluation results including
  - loss curves
  - Confusion Matrix
  - accuracy, sensitivity, specificity, and AUC.

- In your report, explain how you fine-tuned the parameters, provide corresponding evaluation results and your findings.

Submit your Python source codes in a single Python script file and the pre-processed data (CSV format).

PART II [5 MARKS]

In this part, we are going to apply K-means clustering on a set of signal data from a phased array of 16 high-frequency antennas. The data is provided in a file called "ionosphere.csv".

Dataset description:

This radar data was collected by a system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kilowatts. The targets were free electrons in the ionosphere. "Good" (g) radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" (b) returns are those that do not; their signals pass through the ionosphere. Your target attribute is "y"; whereas all other attributes in the dataset with their names starting with "a" are signal readings (features).

Tasks:

a. Cluster the data using the scikit-learn K-means clustering with K = 2 and specifying random_state = 0. Submit your Python script file for this process.

b. Use the clustering results obtained from above task and the class label "y" to count and fill in the number of signals for each of the four categories in the table below.

	y="g"	**y="b"**
Cluster 0		
Cluster 1		

Submit your Python source codes for Part II (a) and (b) in a single Python script file. No marks will be given to your answers unless the relevant source codes are submitted.

PART III [5 MARKS]

Given six data objects (P1, P2,…..,P6), their similarity matrix is shown in the below table. Apply agglomerative hierarchical clustering tree on the data objects **MANUALLY**. Merge the clusters by using **Max distance** and update the **similarity matrix** correspondingly.

	P1	P2	P3	P4	P5	P6
P1	1.00	0.57	0.29	0.29	0.57	0.43
P2	0.57	1.00	0.14	0.43	0.43	0.00
P3	0.29	0.14	1.00	0.71	0.43	0.86
P4	0.29	0.43	0.71	1.00	0.43	0.57
P5	0.57	0.43	0.43	0.43	1.00	0.57
P6	0.43	0.00	0.86	0.57	0.57	1.00

Requirements

- In your report, show each step of the clustering clearly, and provide the final dendrogram diagram. **No marks will be given to your answers unless the step-by-step results are provided.**
- You are welcome to use pen and paper to complete Part 3 and attach photos or scanned copies to your report.

IMPORTANT NOTES

1. A penalty of **5%** of the marks per day will be imposed on late submissions of assessment up to five (5) working days after the due date. **An assignment submitted more than FIVE working days after the due date will NOT be accepted, and ZERO mark will be assigned.**
2. If you need an extension of up to five business days, you should apply using the 'Request for Extension' form **up to three business days before the due date** of your assessment.

   When you cannot apply for a short extension prior to the deadline, or you need more than a five-day extension, you should review the criteria on https://www.latrobe.edu.au/students/admin/forms/special-consideration/eligibility-criteria to see if you are eligible to apply for Special Consideration. More information can be found in Assessment Procedure - Adjustments to Assessment (incorporating Special Consideration) / Document / La Trobe Policy Library.

   Subject coordinator is Dr. Lydia Cui L.Cui@latrobe.edu.au. Please note that submitting wrong forms or providing the wrong contact email will result in delays of assessing your request.

3. Academic misconduct includes poor referencing, plagiarism, copying and cheating. **Copying, Plagiarism**: Plagiarism is the submission of somebody else's work in a manner that gives the impression that the work is your own. Recall that **the University takes academic misconduct very seriously**. When it is detected, penalties are strictly imposed. You should familiarise yourself with your responsibilities about Academic Integrity. Detailed information can be found here: http://www.latrobe.edu.au/students/learning/academic-integrity

SUBMISSION GUIDELINE

- Submit before 16th Oct 2022 11:59 PM (Australian Eastern Standard Time) Sunday (Week 12).
- Upload a single .zip archive onto LMS before the deadline. The .zip archive needs to be named with your student ID (SID), e.g., if your SID is “12345678”, then the archive must be called “12345678.zip”. It should contain
    - A document (word or PDF) of your answers to Part I, II and III. The document needs to be named with your SID, e.g., if your SID is “12345678”, then name the document as “12345678_report.pdf” or “12345678_report.docx” or “12345678_report.doc”
        - Explain each step and show your results for all the questions in the document
    - Python source code to support your answers
    - Pre-processed CSV file.
- Assignment submitted without Python source code will not be evaluated.
- Late submissions will incur a penalty of **5% of the marks per day**.

END