SIT720 Machine Learning
Assessment Task 2: Problem solving task.

This document supplies detailed information on Assessment Task 2 for this unit.

Key information

- Due: **Week 7, Monday 30 August 2021** by 8.00 pm (AEST),
- Weighting: 15%

Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
ULO2 - Perform unsupervised learning of data such as clustering and dimensionality reduction.	**GLO1 -** through the assessment of student ability to use data acquisition techniques to obtain, manipulate and represent data. **GLO3 -** through student ability to use specific programming language and modules to obtain, pre-process, transform and analyse data. **GLO4** -through assessment of student ability to make decisions to obtain data, use appropriate techniques to represent and visualise complex relationships in the data. **GLO5 -** through assessment of student ability to solve problems relates to ill-defined data.

Purpose

This assessment task is for student to apply skills for data clustering and dimensionality reduction. Students will be required to demonstrate ability in data representation, and competency in applying suitable clustering/dimensionality reduction techniques in a real-world scenario.

Assessment 2 **Total marks = 40**

Submission Instructions

a) Submit your solution codes into a **notebook file with ".ipynb"** extension. Write discussions and explanations including outputs and figures into a separate file and **submit as a PDF file**.
b) Submission other than the above-mentioned file formats will not be assessed and given **zero** for the entire submission.
c) Insert your Python code responses into the cell of your submitted ".ipynb" file **followed by the question** i.e., copy the question by adding a cell before the solution cell. If you need multiple cells for better presentation of the code, add question only before the first solution cell.
d) Your submitted code should be executable. If your **code does not generate** the submitted solution, then you will **get zero** for that part of the marks.
e) Answers must be **relevant and precise**.
f) No **hard coding** is allowed. Avoid using specific value that can be calculated from the dataprovided.
g) Use **topics covered till week 6** for answering this assignment.
h) Submit your assignment **after running each cell individually**.
i) The submitted notebook **file name** should be of this form "SIT720_A2_studentID.ipynb". For example, if your student ID is 1234, then the submitted file name should be "SIT720_A2_1234.ipynb".

Questions

Datafile: Download the dataset (.csv) from the SCADI .
Data Description: This dataset contains 206 attributes of 70 children with physical and motor disability based on ICF-CY. For more information click this link.

1. Determine the number of subgroups from the dataset using attributes 3 to 205 i.e., exclude attributes 1, 2 and 206. Is this number same as number of classes presented by attribute 206? Explain and justify your findings. **4 marks**
2. Is this data facing curse of dimensionality? If so, then how to solve this problem. Explain with a two-dimensional plot and report relevant loss of information. **4 marks**
3. After applying principal component analysis (PCA) on a given dataset, it was found that the percentage of variance for the first N components is X%. How is this percentage of variance computed? **2 marks**

Background
Obesity has become a global epidemic that has doubled since 1980, with serious consequences for health in children, teenagers, and adults. Obesity levels in individuals may relate to their eating habits and physical condition. In this assessment, you will be analysing and creating ML models based on a given dataset that contains attributes of individuals with relation to obesity levels.

Dataset filename: obesity_levels.csv

Dataset description: This dataset include data for the estimation of obesity levels in individuals based on their eating habits and physical condition. The data contains 17 attributes and 2111 records.

Features and labels: The attribute names are listed below. The description of the attributes can be found in this article (web-link).

I. Gender
II. Age
III. Height
IV. Weight
V. family_history_with_overweight (family history of overweight)
VI. FAVC (frequent high caloric food)
VII. FCVC (vegetables per meal)
VIII. NCP (number of main meals per day)
IX. CAEC (any food between meals)
X. SMOKE (smoking)
XI. CH2O (daily water intake)
XII. SCC (daily consumed calories)
XIII. FAF (frequency of physical activity)
XIV. TUE (technology usage)
XV. CALC (consumption of alcohol)
XVI. MTRANS (means of transport)
XVII. NObeyesdad (obesity levels, i.e. Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III)

4. Create a machine learning (ML) model for predicting "weight" using all features except "NObeyesdad" and report observed performance. Explain your results based on following criteria: **10 marks**
   a. What model have you selected for solving this problem and why?
   b. Have you made any assumption for the target variable? If so, then why?
   c. What have you done with text variables? Explain.
   d. Have you optimised any model parameters? What is the benefit of this action?
   e. Have you applied any step for handling overfitting or underfitting issue? What is that?

5. Create a ML model for classifying subjects into two classes applying following constraints on above dataset. **12 marks**
   - Use "NObeyesdad" as target variable and rest of them as predictor variables.
   - drop samples with value "Insufficient Weight" for "NObeyesdad"
   - Group Normal Weight, Overweight Level I, and Overweight Level II into a class, and the other three labels (Obesity Type I, II, III) as the other class.

   a. Report classification performance scores. Select scores that you think best for describing the model performance with appropriate justification.
   b. Have you taken any step to check generalisability of the model? What is that and how it ensures generalisability.
   c. Can you design and develop any other model for solving this problem? If so, then why have you used the reported one? Give your justification.

   N. B. Use of multiple models to compare results will increase your chances to get higher marks. This part is for students who are targeting HD – Higher Distinction.

6. Suppose that a company has a number (>=500) of resorts around the globe. **8 marks**
   a. Identify a list of features (>=5) that can be used to describe these resorts.
   b. Create a dataset (rows>=500) and explain all variables. You can generate data either synthetically or collecting from similar datasets. Submit your created dataset. In addition, please provide links in case you have collected the dataset.
   c. Build a ML model that can help a customer to select appropriate set of resorts based on the season of travel. Present and describe the performance of your model.
   d. Why do we need a ML model for this problem?

   N. B. This is a HD (High Distinction) level question. Those students who target HD grade should answer this question (including answering all the above questions). For others, this question is an option. This question aims to demonstrate your expertise in the subject area and the ability to do your own research in the related area.

Submission details

Deakin University has a strict standard on plagiarism as a part of Academic Integrity. To avoid any issues with plagiarism, students are strongly encouraged to run the similarity check with the Turnitin system, which is available through Unistart. A Similarity score MUST NOT exceed 39% in any case. Late submission penalty is 5% per each 24 hours from- Week 7, Monday 30 August 2021 by 8.00 pm (AEST), No marking on any submission after 5 days (24 hours X 5 days from- Week 7, Monday 30 August 2021 by 8.00 pm (AEST)).

Extension requests
Requests for extensions should be made to Unit/Campus Chairs well in advance of the assessment due date. If you wish to seek an extension for an assignment, you will need to submit a request using the "Extension Request" link of the "Assessment" menu in the unit site, as soon as you become aware that you will have difficulty in meeting the scheduled deadline, but at least 3 days before the due date. When you make your request, you must include appropriate documentation (medical certificate, death notice) and a copy of your draft assignment. Conditions under which an extension will normally be approved include:

Medical To cover medical conditions of a serious nature, e.g. hospitalisation, serious injury or chronic illness. Note: Temporary minor ailments such as headaches, colds and minor gastric upsets are not serious medical conditions and are unlikely to be accepted. However, serious cases of these may be considered.

Compassionate e.g. death of close family member, significant family and relationship problems.

Hardship/Trauma e.g. sudden loss or gain of employment, severe disruption to domestic arrangements, victim of crime. Note: Misreading the timetable, exam anxiety or returning home will not be accepted as grounds for consideration.

Special consideration
You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time. See the following link for advice on the application process: http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration.

Assessment feedback
The results with comments will be released within 15 business days from the due date.

Referencing
You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

Academic integrity, plagiarism, and collusion
Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.
Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.
Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.
Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity.