ROBERT GORDON UNIVERSITY ABERDEEN

Assessment Brief - Coursework

Academic Year	**2023-24**
Semester	**2**
Module Number	**CMM 704**
Module Title	**Data Mining**
Assessment Method	**Coursework**
Deadline (time and date)	**Part 1: Report on Assignment: 19th April 2024 Midnight UK time.** **Part 2: Group Presentation: 24th April 2024 to 2nd May 2024**
Submission	**Online via Moodle**
Word Limit	**N/A**
Use of Generative Artificial Intelligence (AI) text	**IS NOT authorised**
Module Co-ordinator	**Dr. Romesh Thanuja**

What knowledge and/or skills will I develop by undertaking the assessment?

To provide an understanding of the main principles underlying Data Mining applied to real-world datasets. To also provide specialised knowledge and valuable insights into algorithms that are at the forefront of machine learning research.

On successful completion of the assessment students will be able to achieve the following Learning Outcomes:

1. Discuss, compare and contrast the advantages and disadvantages of applying a specific data mining technique to a given learning task.

2. Use a toolkit to develop a data mining application tailored to a given learning task.

3. Effectively interpret the results of learning through an understanding of the strengths and limitations of data mining technology and the selection of an appropriate evaluation technique.

4. Demonstrate knowledge of the state-of-the-art in data mining and an awareness of current areas of research.

5. Apply and, where necessary, adapt an appropriate data mining technique to a given problem.

What is expected of me in this assessment?

Task(s) - content

Coursework Description

Part 1: Report on Assignment.

You have the flexibility to choose your dataset/data sets from sources like UCI repository, Kaggle, OpenML, or any other openly accessible domain. The project involves completing the following three tasks that can be implemented on one or more of the real-world datasets you select.

1. Clustering
2. Regression
3. Classification

You will be expected to present each of the datasets you are analysing, identify research questions that can be addressed by your analysis and present relevant existing literature and contrast your results again it. For each dataset, parts (1), (2), and (3) should be completed. Part (4) should be conducted separately for the relevant data set.

1. Introduction
    a) Clearly define the data mining problem and objectives.
    b) Formulates the problem in a way suitable for data mining techniques.

2. Preprocess the dataset as specified in the data mining process.
    a) Handle missing values and outliers if any.
    b) Produce Q-Q plots and histograms of the features and apply the transformations if required.
    c) If it is required, apply suitable feature coding techniques.
    d) Scale and/or standardize the features and produce relevant graphs to show the scaling/ standardizing effect.
    e) If required, apply suitable feature discretization techniques.

3. Perform feature engineering by executing the following task:
    a) Identify significant and independent features using appropriate techniques.
    b) Show how you selected the features using suitable graphs.

What is expected of me in this assessment?

4. Model selection, training, evaluation by performing the following tasks:

    a) For the clustering task:
        i. Justifies the choice of clustering algorithm for the data set.
        ii. Consider and apply alternative algorithms to the data set and explain why they were chosen.
        iii. Using suitable evaluation matrices, compare the applicability of different clustering algorithms on the given dataset.
        iv. Relates the clusters to the original problem and provides actionable insights.

    b) For the classification task:

        i. Justifies the choice of classification algorithm for the data set.
        ii. Consider and apply alternative algorithms to data set and explain why they were chosen.
        iii. Using suitable evaluation matrices, compare the applicability of different classification algorithms on the given dataset.
        iv. Relate classification results to the original problem and provide actionable insights.

    c) For the regression task:

        i. Perform the following regression techniques on the data set to predict the value of a response variable.

            - Linear regression with cross-validation (K=10)
            - Lasso regression with cross-validation (K=10)
            - Ridge regression with cross-validation (K=10)

        ii. Using suitable evaluation matrices, compare the applicability of different regression algorithms on the given dataset.
        iii. Relates regression modelling results to the original problem and provides actionable insights.

What is expected of me in this assessment?

Part 2: Group Presentation

Description: You are required to form groups consisting of three individuals and explore the use of **deep learning models** to address a substantial issue within the business domain associated with Big Data Analytics. Evaluate and compare three state-of-the-art deep learning techniques, analyzing their respective applicability in solving the selected problem.

Task(s) – format

The Assignment Report

Part 1:

- *You are required to formulate solutions for each of the parts (1) through (4) mentioned above for all three different tasks. For each part, provide a descriptive summary with an interpretation of the output obtained after executing each cell. Clearly explain your Python code and specify the outputs produced by the code for the dataset provided in a Jupyter Notebook named SolutionX_IITNumber.ipynb. Here, X represents the task number. Compile three separate notebooks for three different tasks into one ZIP file before submitting.*

Part 2:

- *Group Presentation of the findings. Each group will get 15 minutes to present their findings.*

How will I be graded?	
A number of subgrades will be provided for each criterion on the feedback grid which is specific to the assessment. The overall grade for the assessment will be calculated using the algorithm below.	
A	At least 50% of the subgrades to be at Grade A, at least 75% of the subgrades to be at Grade B or better, and normally 100% of the subgrades to be at Grade C or better.
B	At least 50% of the subgrades to be at Grade B or better, at least 75% of the subgrades to be at Grade C or better, and normally 100% of the subgrades to be at Grade D or better.
C	At least 50% of the subgrades to be at Grade C or better, and at least 75% of the subgrades to be at Grade D or better.
D	At least 50% of the subgrades to be at Grade D or better, and at least 75% of the subgrades to be at Grade E or better.
E	At least 50% of the subgrades to be at Grade E or better.
F	Failing to achieve at least 50% of the subgrades to be at Grade E or better.
NS	Non-submission.

*If the word count is above the specified word limit by more than 10% or the submission contains an excessive use of text within tables, the grade for the submission will be reduced to the next lowest grade.

ROBERT GORDON UNIVERSITY ABERDEEN

Date created: August 2023
Version: 4