School of Computing

Module Code	COSREP / M26538
Module Title	Applied Machine Learning and Data Mining
Module Coordinator Other lecturers	Dr. Alaa Mohasseb alaa.mohasseb@port.ac.uk
Assessment Item number	Item 1
Assessment Title	AMLDM Course work
Date Issued	2021-10-13

Schedule and Deliverables

Deliverable	Value	Format	Deadline / Date	Late deadline ECF deadline
Part 1	10%	One report file (.pdf) and a single .zip file containing the python source codes (upload it to your GitHub repository)	**2022-05-06** 23:00 [GMT/BST]	2022-05-20 23:00 [GMT/BST]
Part 2	50%			
Part 3	40%			

Notes and Advice

- The Extenuating Circumstances procedure is there to support you if you have had any circumstances (problems) that have been serious or significant enough to prevent you from attending, completing or submitting an assessment on time. If you complete an Extenuating Circumstances Form (ECF) for this assessment, it is important that you use the correct module code, item number and deadline (not the late deadline) given above.
- ASDAC are available to any students who disclose a disability or require additional support for their academic studies with a good set of resources on the ASDAC moodle site
- The University takes any form of academic misconduct (such as plagiarism or cheating) seriously, so please make sure your work is your own. Please ensure you adhere to our *Code of Student Behaviour* and watch the video on Plagiarism.
- Any material included in your coursework should be TECFAC 08 Plagiarism fully cited and referenced in **APA 7** format. Detailed advice on referencing is available from the library .
- Any material submitted that does not meet format or submission guidelines, or falls outside of the submission deadline could be subject to a cap on your overall result or disqualification entirely.

- If you need additional assistance, you can ask your personal tutor, student engagement officer ana.baker@port.ac.uk, academic tutor xia.han@port.ac.uk or your lecturers.
- If you are concerned about your mental well-being, please contact our Well-being service.

Part 1 - Questions (10%)

Answer the following questions in no more than 500 words (100 words for each question) your answers must reflect your understanding, knowledge and insights you have gained throughout this module.

1. Assume that you are a data scientist working on a project for a brand of organic food. The brand wants to create a campaign to promote its organic food. To do so, the best times of day to run online ads need to be determined to promote the brand's web store. To achieve this, you need to collect data from the brand's website about when people typically purchase organic food on it. Using the data quality dimensions, how can you ensure that your data is of high quality? (2pt)
2. When building a classification model, which evaluation metrics do you think is more important to evaluate the performance of the model and why? (2pt)
3. Download the 'Students' dataset and answer the following questions: (2pt)
    a. If you need to build a regression model based on the dataset. Which features would you select and what would the objective of the prediction be?
    b. If you need to build a classification model to predict the students' grades, based on the characteristics of the dataset and objective of the classification model, which classification algorithm would you use and why?
4. Let's say, for example, you are a data scientist working on a project for a retail brand that sells various products. The brand wants to segment customers based on purchase history, interests or activity monitoring to create personalised campaigns. How do you think clustering could be used to help the brand achieve this goal? Explain and provide an example. (2pt)
5. Explain some of the ways that entertainment services like Netflix and Spotify could use association rules. (2pt)

Part 2 - Supervised Learning (50%)

Task I: Classification

1. Select any three of the following datasets

Dataset	Link
Heart attack prediction:	https://www.kaggle.com/imnikhilanand/heart-attack-prediction
Bank Loan prediction	https://www.kaggle.com/ninzaami/loan-predication
Weather	https://www.kaggle.com/jsphyg/weather-dataset-rattle-package/download
Mushroom Classification	https://www.kaggle.com/uciml/mushroom-classification
Online shoppers' intention	https://kaggle.com/roshansharma/online-shoppers-intention

2. Use simple descriptive analytics to analyse the data (e.g. all attributes distribution, outliers).
3. Use data exploratory techniques (e.g. three visualisations) to explore the dataset and analyse the results.
4. Build a classification model for each dataset using three of the following classification techniques.
    - Decision tree
    - RandomForest
    - Support Vector Machine
    - Naive Bayes
    - K-Nearest Neighbors.
5. Analyse the results and compare the performance of the applied algorithms in terms of accuracy.

Task II: Regression

1. Select any two of the following datasets

Dataset	Link
Facebook metrics	https://archive.ics.uci.edu/ml/datasets/Facebook+metrics
Fertility	https://archive.ics.uci.edu/ml/datasets/Fertility
Air Quality	https://archive.ics.uci.edu/ml/datasets/Air+Quality
Energy efficiency	https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

2. Use simple descriptive analytics to analyse the data (e.g. all attributes distribution, outliers).
3. Use data exploratory techniques (e.g. three visualisations) to explore the dataset.
4. Build a regression model for each dataset using the following regression techniques.
    - Linear Regression.
    - Multiple Linear Regression
5. Analyse the results and compare the performance of the applied algorithms.

Part 3 - Unsupervised Learning (40%)

Task I: Clustering

1. Select any two of the following datasets

Dataset	Link
Diabetes	https://www.kaggle.com/faysalislam/diabetes
Absenteeism at work	https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
Credit Card	https://www.kaggle.com/arjunbhasin2013/ccdata
clickstream data for online shopping	https://archive.ics.uci.edu/ml/datasets/clickstream+data+for+online+shopping

2. Use simple descriptive analytics to analyse the data (e.g. all attributes distribution, outliers).
3. Use data exploratory techniques (e.g. three visualisations) to explore the dataset and analyse the results.
4. Build a clustering model for each dataset using two of the following clustering techniques.
    - K-means
    - DBSCAN
    - Hierarchical clustering
5. Analyse the results and compare the performance of the applied algorithms. The clustering analysis could be done on a small subset of the attributes. You need to justify your attribute selection.

Task II: Association Rule

1. Select one of the following datasets

Dataset	Link
Supermarket	https://www.kaggle.com/irfanasrullah/groceries
Heart Failure	https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records

2. Use simple descriptive analytics to analyse the data (e.g. all attributes distribution, outliers).
3. Use data exploratory techniques (e.g. three visualisations) to explore the dataset and analyse the results.
4. Perform association rule mining using only one of the following algorithms and discuss the results.
    - Apriori
    - FP-Growth

Deliverables of the components of the coursework are

- A report documenting the following:
- Part 1 in no more than 500 words reflects your understanding and the insights you have gained throughout this module.
- Part 2 and Part 3 in no more than 2500 words excluding figures and tables and must cover the following areas:
    - A short summary of the machine learning applications and datasets you used and a justification of the chosen algorithms and attributes.
    - A detailed analysis of your results when comparing the different classification techniques.
    - A detailed analysis of your results when comparing the different regression techniques.
    - A detailed analysis of your results when comparing the different clustering techniques.
    - A detailed analysis of your results when reporting the association rule results.

The submission is online through Moodle (the submission details will be available on Moodle).

Please ensure that your coursework is anonymous. Your NAME must not appear anywhere on the coursework or the cover sheet. Please use your ID only

Grading Criteria for Part 2 and part 3

The grading criteria (in 100% breakup of marks) will be applied to your work. The marking scheme below is to help you understand the grading criteria and provide you with a frame of reference for your effort.

Marking Scheme

60% - Analysis of the results of the experiments you have conducted

- Include your own explanation of the chosen problem.
- Provide a clear description of the data.
- Include important findings.
- Critically evaluate the advantages and disadvantages of the algorithms selected.
- Discuss the effect and impact of the algorithms used, justifying your approach and methods.

20% - Appropriate use of tables and figures when reporting the results

- Include clear, informative titles.
- Provide meaningful and concise visualisations.
- Present information using the appropriate visualisations.
- Pay attention to consistency between values or details in a table/figure.
- Use labels and legends to ensure your figures are clear

10% - Conclusion with recommendations on how to match a dataset to a technique

- Include a thorough evaluation of the work you have done and the outcomes of the project.
- Base your conclusions on critical analysis and the insights you have gained throughout this module.

10% - Structure and presentation

- Organise your report so it is clear and easy to read, using sections and subsections as appropriate.
- Pay attention to the quality of your prose, including the use and flow of language, grammar, spelling, and format.
- Use graphics and tables as appropriate and choose a page layout that is easy to follow and understand.