# Assessment Information/Brief 2022-23

To be used for all types of assessment and provided to students at the start of the module. Information provided should be compatible with the detail contained in the approved module specification although may contain more information for clarity.

| | |
|---|---|
| **Module title** | **Machine Learning and Data Mining** |
| **CRN** | 40939 |
| **Level** | 7 |
| **Assessment title** | **Machine Learning & Data Mining with Python and Azure Machine Learning Studio** |
| **Weighting within module** | This assessment is worth 100% of the overall module mark. |
| **Module Leader/Assessment set by** | **Professor Mo Saraee** |
| **Submission deadline date and time** | **4pm 12nd December 2022**<br><br>Students with a Reasonable Adjustment Plan (RAP) or Carer Support Plan should check your plan to see if an extension to this submission date has been agreed. |
| **How to submit** | Your assignment should be submitted through the link provided on Blackboard and should be separated into two formats. First, a **single** PDF report named "Your_Name.pdf" and second, a zipped file containing your codes also named "Your_Name.zip". This should contain working Python code for all parts of the assessment, in either .py or .pynb format. You should also include a clearly written description of each file and its use in a "Read Me.txt" file in the same zip file. |
| **Assessment task details and instructions** | This coursework will give you the opportunity to use the machine learning and data mining techniques covered in this module to analyse datasets that interest you, to draw conclusions based on your analysis, and finally to present your results in the form of a report. |

There are four tasks to this assignment. For each assignment, you should choose a dataset amenable to the data mining task. For example, for Task 1 you should choose a dataset which contains a potential target variable suitable for classification. **You do not have to use the same dataset for all four tasks.**

For each task, you must fully explain and document your full experimental process, including the exploratory data analysis, data preparation and cleaning and the algorithms selected. A writing frame is provided on page six of this document, and you should complete a report for each of the four tasks (please provide all four reports in one PDF document). Overall, your submission should be around 5,500 words excluding references, so each report is likely to be around 1,250-1,500 words.

A report framework is provided at the end of this brief to provide more guidance on the expected contents of the report.

**Task 1 (35 Marks)**

**a)** Apply **two** classification algorithms of your choice on your chosen dataset using Python. Compare the performance of the two algorithms, justifying your choice of performance metrics. You should critically evaluate your classification models, and recommend which, if any, would be appropriate for future deployment. ***(15 Marks)***

**b)** Use Azure Machine Learning Designer to apply classification to the same dataset as you used for part a). ***(15 Marks)***

**An addition 5 marks is available if one of your chosen classification algorithms for part a) is a neural network.**

**Task 2 (20 Marks)**

Apply association rules mining on a selected dataset of your choice using Python. You should provide an analysis and evaluation of the association rules identified, using appropriate metrics to assess their value.

**Task 3 (25 Marks)**

Apply **two** clustering algorithms on a selected dataset of your choice using Python. You should provide an analysis and evaluation of the

clusters identified and discuss which clustering method may be better suited to your data.

**Task 4 (20 Marks)**

Using Python, apply text mining and sentiment analysis on 30 hotels or restaurants from the **tourist_accommodation_reviews.**csv dataset accompanying this brief. You can select any 30 hotels from the data, but you should provide a logical reason for your selection (e.g., based on the region).

| | |
|---|---|
| **Assessment Criteria** | Assessment criteria are provided alongside this brief.<br><br>You should look at the assessment criteria to find out what we are specifically looking at during the assessment. |
| **Knowledge and Understanding** | **Assessed intended learning outcomes**<br>On successful completion of this assessment, you will be able to:<br><br>1. Critically assess diverse issues regarding the use of data mining and machine learning in real-world contexts, including ethics<br>2. Design, build and use business intelligence systems, justifying decisions made<br>3. Design and create reports to present analytical and interpretative information in creative and effective ways<br>4. Devise strategies for making effective use of analytical software such as Python and Azure Machine Learning Studio.<br>5. Learn about different algorithms, such as classification and clustering. |
| **Practical, Professional or Subject Specific Skills** | 1. Diverse issues regarding the use of data mining techniques to real-world datasets<br>2. Discover patterns within a dataset using exploratory data analysis<br>3. Use of Python / Azure for data mining<br>4. Discover techniques to leverage Python's features and work with its libraries<br>5. Reporting and presentation of analytical and interpretative information |

| | |
|---|---|
| **Employability Skills developed / demonstrated** | **Communication**<br>**Critical Thinking and Problem Solving**<br>**Data Literacy**<br>**Digital Literacy**<br>**Industry Awareness**<br>**Innovation and Creativity**<br>**Self-management and Organisation** |
| **Word count/ duration (if applicable)** | Your assessment should be 5,500-8,000 words in total across the four tasks. |
| **Academic Integrity and Referencing** | Students are expected to learn and demonstrate skills associated with good academic conduct (academic integrity). Good academic conduct includes the use of clear and correct referencing of source materials. Here is a link to where you can find out more about the skills which students need:<br>Academic integrity & referencing<br>Referencing<br><br>**Academic Misconduct is an action which may give you an unfair advantage in your academic work. This includes plagiarism, asking someone else to write your assessment for you or taking notes into an exam. The University takes all forms of academic misconduct seriously.** |
| **Assessment Information and Support** | **Support for this Assessment**<br>You can obtain support for this assessment by contacting Prof. Mo Saraee or attending one of our drop-in sessions. Details of these will be released on Blackboard.<br><br>You can find more information about understanding your assessment brief and assessment tips for success here.<br><br>**Assessment Rules and Processes**<br>You can find information about assessment rules and processes in Blackboard in the Assessment Support module.<br><br>**Develop your Academic and Digital Skills**<br>Find resources to help you develop your skills here.<br><br>**Concerns about Studies or Progress** |

| | |
|---|---|
| | If you have any concerns about your studies, contact your Academic Progress Review Tutor/Personal Tutor or your Student Progression Administrator (SPA).<br><br>**askUS Services**<br>The University offers a range of support services for students through askUS including Disability and Learner Support, Wellbeing and Counselling Services.<br><br>**Personal Mitigating Circumstances (PMCs)**<br>If personal mitigating circumstances (e.g. illness or other personal circumstances) may have affected your ability to complete this assessment, you can find more information about the Personal Mitigating Circumstances Procedure here. Independent advice is available from the Students' Union Advice Centre about this process. Click here for an appointment to speak to an adviser or email advicecentre-ussu@salford.ac.uk. |
| **In Year Retrieval Scheme** | Your assessment is/is not (please delete as appropriate) eligible for in year retrieval. If you are eligible for this scheme, you will be contacted shortly after the feedback deadline.<br><br>You can find more information about this scheme in Blackboard in the Assessment Support module. |
| **Reassessment** | For students with accepted personal mitigating circumstances for absence/non submission, this will be your replacement assessment attempt.<br><br>We know that having to undergo a reassessment can be challenging however support is available. Have a look at all the sources of support outlined earlier in this brief and refer to the Personal Effectiveness resources. |

## Key elements of the report

### Title

The title should provide an overview of the focus of your problem and the expected solution.

### Introduction

This section contains a brief background to the topic and leads to the formulation of the specific question, based on your selected topic. The research question must be focused and clear.

### Datasets

You are welcome to choose any datasets that interest you, and that has enough data to enable meaningful analysis. In making your choice, you should be sure to consider what problem or problems you would be able to solve by employing data mining on the dataset. In other words, you should ask yourself: How could I use data mining to answer one or more questions about the datasets?

### Explanation and preparation of datasets

Briefly describe the datasets you have used, independent and dependent variables. Explain any preparation tasks (e.g., normalisation, dealing with missing values, handling class imbalance etc.) carried on the datasets.

### Implementation in Python / Azure Machine Learning Designer

Implement your proposed approach using libraries available in Python. This section will include:

- A brief description of the algorithms used.
- The application of data-mining techniques to selected datasets that you choose using Python (or Azure Machine Learning Designer for Task 1b).
- Explanation of the experimental procedure, including the setting and optimisation of model hyperparameters during training, and your approach to validation (for supervised learning tasks).
- Visualisation of the results.

**Results analysis and discussion**

- Explain and justify the performance metric you choose to use to evaluate the model(s).
- A clear and compelling presentation of the results that you obtain, both from the data mining and any other analysis that you may perform.
- For tasks that require you to use more than one algorithm, you should compare and discuss the results obtained from each.
- You should also consider and discuss any ethical, legal or professional considerations in using machine learning and data mining on the datasets you have selected.

**Conclusions**

The key points from the assignment must be synthesised within the conclusion. This must relate back to the introduction and the research question and provide an overall evaluation of the validity of the solution you have proposed.

**References**

You will list all publications referenced in the report. You should show evidence of sufficient readings related to your work. References must follow the Harvard formatting system as in this guide:

http://www.salford.ac.uk/library/help/user-guides/general/Bibliographic-Citations-APA-QuickRef-Apr2015.pdf

**Appendices**

Appendices may be used to provide relevant supporting evidence for reference but should only be used if necessary. Students may wish to include in appendices, evidence which confirms the originality of their work or illustrates points of principle set out in the main text.

**Sample datasets**

These links are provided as examples of datasets which may be appropriate to the given task. You are welcome to use them if you wish, but we also encourage you to research and identify your own datasets for these tasks. You may want to choose ones in domains you have existing experience or interest in.

**Classification**

https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

https://www.openintro.org/stat/data/?data=ames

https://archive.ics.uci.edu/ml/datasets/adult

https://data.world/data-society/pima-indians-diabetes-database

https://archive.ics.uci.edu/ml/datasets/car+evaluation

**Association Rule Mining**

https://archive.ics.uci.edu/ml/datasets/online+retail

https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

https://catalog.data.gov/dataset/association-rule-mining-data-for-census-tract-chemical-exposure-analysis

http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data

https://catalog.data.gov/dataset/2011-american-community-survey-1-year-pums-person-file

**Clustering**

https://www.openintro.org/stat/data/?data=gun_violence

https://catalog.data.gov/dataset/association-rule-mining-data-for-census-tract-chemical-exposure-analysis

https://catalog.data.gov/dataset/asthma-hospitalization-and-ed-visit-primary-diagnosis- crude-and-age-adjusted-rates-by-town

http://hdr.undp.org/en/data

https://catalog.data.gov/dataset/2011-american-community-survey-1-year-pums-person-file

## Assessment criteria

| **Overall level** | 0-19% | | 20-39% | | 40-59% | | 60-79% | | 80-100% | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Extremely poor | Very poor | Poor | Inadequate | Unsatisfactory | Satisfactory | Good | Very Good | Excellent | Outstanding |
| **Title, Introduction, Conclusion (20%)** | No Title/Very vague title | | Uninformative title, vague introduction, unreliable conclusion | | Satisfactory title, introduction well defines the studied problem and the intended tasks, clear conclusion | | Informative and attractive title, clear setting of the scene and boundaries of the report in introduction, conclusion drawn persuasively from results analysis and discussion. | | Concise and appealing title, introduction presents an excellent clarity of focus of the report, conclusions are reliable and can be trustfully used by users. | |
| **Explanation of datasets, legal and ethical issues (if any) and References (10%)** | Did not perform data preparation steps for ML | | Insufficient collection of primary information, datasets are barely explained. Inadequate attempt made at proper referencing, many errors/omissions<br><br>Direct download and no preparation of data for data mining task. | | Adequate engagement with relevant information collection, reasonable fraction from primary sources. Adequate dataset explanation. Acceptable attempt made at proper referencing, with a number of errors/omissions. | | Good information collection, relevant to the assignment, significant fraction from primary sources. Datasets clearly explained. Referencing good, but with some errors and omissions.<br><br>Detailed handling of data from different sources also considered ethical and legal issues. | | Information collection of very high standard, relevant to assignment and mostly from primary sources.<br>Concise and informative dataset explanation.<br>Referencing almost perfect.<br><br>Outstanding handling of data from different sources also considered important legal and ethical issues. | |

| | | | | | |
|---|---|---|---|---|---|
| **Implementation in Python (or Azure Machine Learning Designer for Task 1b) (40%)** | No implementation<br><br>Implementation not justified for the task considered. | Experimental implementation and setup is lacking detail, little or no relevant description and discussion of relevant package and functions, and no critique of designs. | Basic descriptions of experiments, design, and statistics that could be conducted, relevant literature is lacking, little or no critique. | Good descriptions of experiments, design, statistics that could be conducted, some relevant literature and basic critique.<br><br>Detailed justifying of decision made for ethical principles throughout the data mining algorithmic selection and usage. | Detailed descriptions of experiments, design, statistics that could be conducted, relevant literature and critique.<br><br>Outstanding justifying of decision made for ethical principles throughout the design, build and use of business intelligence systems and data mining algorithmic selection.<br>of decision made for |
| **Results analysis and discussion (30%)** | No results interpretation or discussion was presented in the report | Results are not presented professionally, little or no results analysis and discussion | Results are presented using proper means such as tables and graphs, results analysis and discussion is general and shallow. | Results are clearly and informatively presented. Results analysis and discussion are specific and sufficient. | Results are professionally presented at standard of a journal publication. Results are critically analysed and discussed. Valuable observation and finding are made from the results. |