University of Roehampton London

School of Arts, Humanities and Social Science

Module title and code: Applications of Data Science (CMP020L014)
Title of coursework: Coursework (Portfolio)

Learning outcomes:	• LO1: Demonstrate a comprehensive understanding of current developments in data science. • LO2: Systematically and critically analyse and evaluate diverse sources of data to solve a problem. • LO3: Propose and develop a data science solution for a complex dataset.
Assessment weighting	Mid-term presentation: 25%, Final report: 75%
Maximum mark	100
Submission details (e.g. submission link)	Part 1: Mid-term presentation: https://moodle.roehampton.ac.uk/mod/assign/view.php?id=1676194 Part 2: Final report: https://moodle.roehampton.ac.uk/mod/assign/view.php?id=1688198
Word limit (if applicable)	4000-word limit for a final report.
Date set	• COVID-19 X-ray: https://www.kaggle.com/datasets/alifrahman/chestxraydataset • Wet signature: https://www.kaggle.com/datasets/saurabstha5/signature-forgery-dataset/data • Plant disease: https://www.kaggle.com/datasets/dhamur/cotton-plant-disease • Low light: https://www.kaggle.com/datasets/soumikrakshit/dark-face-dataset • Underwater: https://www.kaggle.com/datasets/vencerlanz09/deep-fish-object-detection
Deadline	Mid-term presentation: 28/03/2024 Final report: 29/04/2024
Feedback and marks	Please see the rubric at the end of this coursework
Assessment setter's name	Mohammad F Khan

ACADEMIC MISCONDUCT:

"Academic integrity and honesty are fundamental to the academic work you produce at the University of Roehampton. You are expected to complete coursework which is your own and which is referenced appropriately. The university has in place measures to detect academic dishonesty in all its forms. If you are found to be cheating or attempting to gain an unfair advantage over other students in any way, this is considered academic misconduct, and you will be penalised accordingly. "

Further details about "Student Code of Conduct" and "Disciplinary Regulations" can be found at: https://www.roehampton.ac.uk/corporate-information/policies/

I. ASSESSMENT INTRODUCTION (IF APPLICABLE):

This group assignment is designed to test your skills to propose and develop a data science solution for a complex dataset, along with your ability to present the findings in a report.
For this group coursework, you are required to carry out a set of tasks by selecting one problem from the list given in Task 1, applying various techniques such as feature engineering/extraction, data acquisition, visualisation, and statistics to develop a refined machine learning algorithms, which includes comprehensive evaluation and analyses of results. You are expected to utilise tools that are relevant to the field of data science.
In this coursework, you should not use neural networks/deep learning as a tool for any type of analysis but can opt for the article that uses it. You are expected to develop an alternate machine-learning algorithm for the chosen problem, which will demonstrate your comprehensive understanding of developing diverse solutions in data science.

- ✓ Task 1: For this coursework, you are required to select one of the following real-world problems and work as a group to investigate a possible solution(s). Alternatively, you can identify a similar problem in a different domain.
  Problem 1: https://www.sciencedirect.com/science/article/pii/S1746809422000520
  Problem 2: https://www.mdpi.com/2076-3417/10/11/3716
  Problem 3: https://link.springer.com/article/10.1007/s11042-022-13598-1
  Problem 4: https://link.springer.com/article/10.1007/s11042-022-12518-7
  Problem 5: https://ieeexplore.ieee.org/document/10167628
  Refer to the following section III-A, to download the articles mentioned above. If required, you can alternatively opt for a similar type of problem having a different application domain with a different dataset, which must follow Tasks 2-5 and be decided after a detailed discussion with the module tutor. Refer to section III-B guidelines to decide on a new problem.
- ✓ Task 2: The problem should contain the mathematical concept of image/video pre-processing, such as feature engineering.
- ✓ Task 3: The dataset should contain at least 100 images or 10 video samples.
- ✓ Task 4: Results and discussion should include tools from descriptive and inferential statistics.
- ✓ Task 5: The final solution should contain a machine learning (except neural networks) algorithm as a part of the analysis.

II. DELIVERABLES (WHAT YOU WILL NEED TO SUBMIT):

The coursework is comprised of two parts, and you will be required to work with your group to present the work in the seminar and prepare a final written report.

1. Present your mid-term presentation in the seminar and submit the presentation slides using the link given on the first page under the submission details heading mentioning Mid-term presentation. Your presentation must have 10-15 presentation slides. The presentation slides must include:

- The state-of-art literature related to the problem you have opted for.
- Explaining the data by using tools from data visualisation.
- Present the statistical analysis of the data you have opted for.
- Discuss the part of the solution you have implemented to solve that problem.
- Present a vision of how you are going to apply machine learning to that problem.

The seminar is like a Q&A session, where other students may ask questions to the presenting group.

2. To submit the final coursework, use the template given on the module Moodle page in the assessment section, and submit the coursework with MATLAB (.m) file using the link given on the first page under the submission details heading mentioning the Final report. Note that, you are expected to code in MATLAB by using your knowledge of fundamental programming demonstrating your understanding to develop a data science solution for a complex dataset.

III. ADDITIONAL INFORMATION (IF REQUIRED):

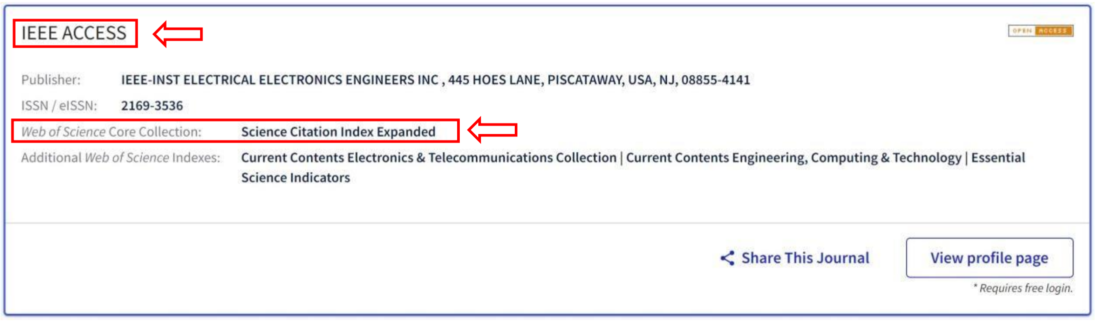
A. HOW TO DOWNLOAD RESTRICTED PAPERS MENTIONED IN THE ASSESSMENT INTRODUCTION:

1. For the ScienceDirect paper given in Problem 1, visit: https://library.roehampton.ac.uk/sciencedirect, and for the Springer paper given in Problems 2 and 4, visit: https://library.roehampton.ac.uk/springercompact
2. Use your university credentials to log in, search for the paper title, and download the paper.

B. HOW TO DECIDE A NEW PROBLEM:

Decide the topic you are willing to pick that should belong to the Science Citation Index Expanded (SCIE) and involve images/videos as a dataset. Data can be collected by using personal devices or downloaded from online open sources. The process for deciding a problem can be conducted by using the following steps:

- Search the topic in Google Scholar (https://scholar.google.com/).
- Look for the related articles in the search result that should not be older than 5 years.
- Define the modification you are planning in the reference article.
- Confirm if the article of your choice falls in the SCIE Journal category by searching the journal name in MJL (https://mjl.clarivate.com/search-results) and appears in the Web of Science Core Collection: Science Citation Index Expanded list. In case you have any confusion, send me the name of the paper and its complete citation details, and I will cross-verify it for you.
- Refer to the screenshot below from MJL illustrating the example paper mentioned in Problem 5 (https://ieeexplore.ieee.org/document/10167628) which belongs to IEEE Access journal:



IEEE ACCESS

Publisher: IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC , 445 HOES LANE, PISCATAWAY, USA, NJ, 08855-4141

ISSN / eISSN: 2169-3536

Web of Science Core Collection: Science Citation Index Expanded

Additional *Web of Science* Indexes: Current Contents Electronics & Telecommunications Collection | Current Contents Engineering, Computing & Technology | Essential Science Indicators

C. DATA COLLECTION PRIVACY

If you are planning to collect the data yourself, then it should not violate the privacy of anyone in any form. Refer: https://ico.org.uk/for-organisations/direct-marketing-and-privacy-and-electronic-communications/guidance-for-the-use-of-personal-data-in-political-campaigning-1/collecting-personal-data

IV. SIMILARITY AND PLAGIARISM

The final submitted report should have the lowest possible similarity percentage.
Currently UoR accepts maximum 20% similarity for research degrees (such as PhD): https://www.roehampton.ac.uk/globalassets/documents/graduate-school/current-students/research-degrees-handbook-2022-23.pdf

Refer the link to understand more about similarity and plagiarism:
https://roehamptonlearning.com/eLearningServices/?p=4010

V. STUDIOSITY

Prior to submitting your final coursework, you are encouraged to submit your draft report to studiosity, make recommended changes to the report, and submit it as final coursework to Moodle for marking.
Refer:
https://www.studiosity.com/service/access?utm_referrer=https%3A%2F%2Fwww.studiosity.com%2F

VI. ASSESSMENT EXPECTATIONS AND RUBRIC:

	Criteria	Expectation	Maximum marks (100)
Presentation (in class)	Mid-term presentation (max. 15 minutes)	The state-of-art literature related to the problem you have opted. Explaining the data by using tools from data visualisation. Present the statistical analysis of the data you have opted. Discuss the part of the solution you have implemented to solve that problem. Present a vision of how you are going to apply machine learning to that problem.	25
Final report	Abstract, conclusion and format demonstration	A brief 200-300 word glance of the problem statement and its possible solution along with results. Demonstrating report by using appropriate language, clear formatting and correct referencing.	10
	Introduction/Literature review appropriateness	Detailed survey of related work that covers state-of-art algorithms on the chosen problem. Defining the modifications, you have conducted in the reference article.	10
	Mathematical understanding and feature engineering	Detailed explanation of algorithmic equations used in the study. Also, showing ability to define a part of a problem in the scope of algebra, calculus, probability, approximation theory and/or numerical analysis.	20
	Statistical analysis	Appropriate and detailed inferential and descriptive statistical analyses conducted on the dataset to refine the possible solution.	10
	Data visualisation	Various types of graphical representation attempted to visualise the dataset as well as simulation results. Also showing ability to efficiently visualising overlapping complex data distribution/simulation results in single plots.	5
	Application of machine learning algorithm	Multiple algorithms have been used for comparison purpose. and the comprehensive analysis has been conducted with detailed reasoning.	10
	Programming language used/Statistical software used	A clear MATLAB code has been developed without using inbuilt functions with proper comments.	10

VII. ASSESSMENT PROCESS AND CRITERIA (PASSING GRADE BOUNDARIES):

52-58%:	Pass work demonstrates: - acceptable attainment of learning outcomes. - evidence of some critical evaluation but little originality. - adequate coverage and understanding of material largely based on teaching material or core texts. - some errors or omissions. - evidence of a basic argument, but insufficiently supported or developed. - adequate writing, with some problems with organisation.
62-68%:	Merit work demonstrates: - convincing attainment of learning outcomes. - evidence of critical evaluation and development of an independent argument. - comprehensive up-to-date range of relevant theoretical and empirical material showing wide reading. - clear understanding of key concepts. - generally good writing and structure.
72-78%:	Distinction work in this range demonstrates: - exemplary attainment of learning outcomes. - a high level of insight and critical evaluation of the material. - a comprehensive and up-to-date account of relevant theoretical and empirical material at the forefront of the discipline. - a thorough understanding and integration of material supporting a cogent argument. - excellent writing of a high academic standard.
82-88%:	Distinction work in this range demonstrates: - attainment beyond the intended learning outcomes. - an outstanding level of originality and creativity, providing a significant new perspective on the question or topic. - a clear, elegant and well supported argument, based on the integration and sophisticated critical evaluation of a substantial body of knowledge. - suitability for publication in a high-quality journal.
100%:	An assessment that could not be bettered within the time available.