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 report: 25%, Final report: 75%
Maximum mark	100
Submission details (e.g. submission link)	**Part 1:** Submit the mid-term presentation report in PDF format: https://moodle.roehampton.ac.uk/mod/assign/view.php?id=2097536 **Part 2:** Submit the final report in PDF format and MATLAB code in .m/.mlx format: https://moodle.roehampton.ac.uk/mod/assign/view.php?id=2097537
Word limit (if applicable)	2500-word limit for a mid-term presentation report. 4000-word limit for a final report.
Date set	24/02/2025
Deadline	**Mid-term presentation report:** 30/03/2025 **Final report:** 14/04/2025
Feedback and marks	For feedback, refer to the rubric at the end of this coursework. Marks will be released within 4 weeks of the final submission deadline.
Assessment setter’s name	Dr Mohammad F Khan

1) ASSESSMENT OVERVIEW:

This assignment evaluates your skills in proposing and developing a data science solution for a complex dataset and presenting your findings in a report. You are required to complete a set of tasks by selecting one problem from the options in **Task 1** (or choosing a similar problem as per Appendix 2), applying techniques such as feature engineering, visualisation, and statistical analysis to enhance machine learning algorithms, and comprehensively analysing and reporting your results.

Important: Do not use neural networks or deep learning in this coursework. If you choose a research article that uses neural networks, replace that section with an alternative explainable AI algorithm.

2) 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/

3) TASKS:

- **Task 1:** Select one of the following real-world problems to investigate and solve. Alternatively, you may identify a similar problem in a different domain (see Appendix 2).

  - **Problem 1: Classification of diabetic retinopathy**
    Diabetic retinopathy is a leading cause of vision impairment and blindness among diabetic patients. Early detection through retinal image analysis is crucial for timely intervention. However, manual diagnosis by ophthalmologists is time-consuming, subjective, and prone to variability. Automated classification of diabetic retinopathy using machine learning can improve accuracy and efficiency, but challenges such as imbalanced datasets, feature extraction, and differentiation between severity levels persist. Developing a robust classification model that effectively distinguishes diabetic retinopathy stages from retinal fundus images is essential for enhancing diagnostic reliability and patient outcomes.

    Resources:
    - Research article: https://www.mdpi.com/2075-4418/12/9/2262
    - Dataset: https://zenodo.org/records/4647952#.YGNjXVUzbIU

  - **Problem 2: Detecting wet signature forgery through images**
    Detecting forgery in wet signatures through images is essential to safeguard the authenticity of legal and financial documents. Wet signatures are widely used for verifying identity and consent; forgery can lead to fraud, data breaches, and unauthorised transactions. Image-based detection allows for detailed analysis of signature characteristics, such as stroke patterns, pressure points, and ink flow, which are difficult to replicate perfectly. Advanced techniques like AI and machine learning enhance accuracy, identifying subtle inconsistencies that might escape human observation. This ensures the integrity of documents, protects individuals and organisations from potential fraud, and upholds trust in sensitive transactions.

    Resources:

- Research article: https://www.mdpi.com/2076-3417/10/11/3716
- Dataset: https://www.kaggle.com/datasets/akashgundu/signature-verification-dataset

➢ **Problem 3: Detecting wildfire using surveillance data**

Wildfires pose a significant threat to ecosystems, human lives, and infrastructure. Early detection is crucial for effective mitigation, but traditional methods relying on satellite imagery and ground-based sensors often suffer from delays and limited coverage. Surveillance video data offers a real-time alternative for wildfire detection. Still, challenges such as varying lighting conditions, smoke interference, and false alarms due to similar visual patterns make accurate identification difficult. Hence it is necessary to develop an advanced machine learning-based model to detect wildfires in surveillance videos, enhancing early warning systems and enabling rapid response to minimise environmental and economic damage.

Resources:

- Research article: https://ieeexplore.ieee.org/abstract/document/9690875
- Dataset: https://datasets.omdena.com/dataset/onfire-dataset

Note: Use a university PC for downloading the aforementioned research articles or refer to Appendix 1 for downloading the research articles at home by using a university library login. 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 2-7 given below. The new problem must be decided after a detailed discussion with the module tutor. For more information, refer to the guidelines given in Appendix 2 for deciding on a new problem.

- **Task 2:** Refine the dataset by choosing at least 100 images or 01 video sample (3-second video recorded at ≥30 fps) for each class.

  Helpful tip: Choosing the image/video dataset for binary classification involves several key considerations. First, ensure the dataset has balanced classes to avoid bias in model training. The images should be relevant to the classification task and of sufficient quality, with clear features distinguishing the two categories. Consider the size of the dataset; larger datasets provide more training examples, improving model generalisation. Additionally, check for labelled data to facilitate supervised learning. Dataset diversity, including variations in lighting, angles, and backgrounds, is crucial for building a robust model.

- **Task 3:** Formulate a mathematical concept for feature engineering for the image/video data.

  Helpful tip: Formulating a mathematical concept for feature engineering in the data involves designing and defining transformations that extract meaningful patterns. To reveal the patterns that are not visible in the spatial domain, you can think of starting by representing images as matrices of pixel intensities; applying pre-processing techniques like mathematical filters to detect edge and texture-based features in the data; reducing the dimensionality of the data through principal component analysis (PCA); projecting the data into lower-dimensional spaces while retaining key characteristics; applying advanced feature engineering methods like HoG transform, Fourier transform, wavelet transform et cetera to analyse the frequency components. The goal of this task is to encode complex information into compact, informative features that can enhance model performance and enable accurate classification.

- **Task 4:** Apply descriptive and inferential statistical tools to analyse the data and test the model performance.

Helpful tip: Descriptive statistics summarise image data by calculating metrics like mean, median, kurtosis, skewness, standard deviation, and pixel intensity distributions to understand patterns and variability. They can help in identifying data imbalances or anomalies before model training. For inferential statistics, techniques like hypothesis testing and confidence intervals evaluate relationships in image features, such as comparing pixel intensity distributions across classes. In machine learning classification, these statistics assess feature relevance, help refine pre-processing steps, and validate model assumptions. Inferential statistics also help test the significance of model performance metrics, ensuring robust conclusions about the classifier's effectiveness on unseen image data. Together, they may enhance the data-driven decision-making process.

- **Task 5:** Create appropriate visualisations, e.g. 2D/3D plots of the complex data and simulation results.

  Helpful tip: Employ colour, size, transparency, and marker shapes to encode additional variables in the plot without overcrowding the plot. Use different scaling functions for different feature values, and visualise the scale in the single plot to maintain the interpretability and clarity of the data.

- **Task 6:** Incorporate machine learning algorithms (excluding neural networks) in your final solution.

  Helpful tip: To apply machine learning algorithms to features extracted from image data for binary classification, start by looking into the data obtained from the feature engineering step, ensuring it is normalised for optimal machine learning performance. Feed these features into the machine learning model. Use hyper-optimisation to handle non-linearly separable data. Evaluate the model using metrics like accuracy, sensitivity, specificity, recall, F1-score, AUC et cetera to ensure robust classification performance.

- **Task 7:** Develop clear, well-commented MATLAB code without relying on built-in functions. The code must be executable on a university machine that can load the cloud dataset and produce the reported output without issues.

  Helpful tip: Writing good comments in MATLAB code enhances readability and helps others understand the logic. Begin by commenting at the start of the script or function, explaining its purpose, inputs, and outputs. Use inline comments to clarify complex or non-obvious code, explaining the reasoning behind key steps or formulas. Keep comments concise, but informative - avoid redundancy, and focus on what the code does, rather than how. Group-related blocks of code with section comments for better organisation. Ensure comments are up-to-date and relevant as code evolves, improving maintainability and making it easier for collaborators or future you to understand the code's functionality.

4) DELIVERABLES (WHAT YOU WILL NEED TO SUBMIT):

Submit your work in two parts:

- **Part 1: Mid-term Presentation and Report (25 marks)**
  Present your progress in a 10-minute seminar talk with 10-15 slides, followed by a 5-minute Q&A. The presentation slides and 2500-word mid-term report (in PDF format) should include the following information:
  - ➢ The literature review related to the problem you have opted for from Task 1.
  - ➢ Explanation of the dataset by using tools from data visualisation.
  - ➢ Descriptive statistical analysis of the dataset you have opted for.

- Discussion on the part of the solution you have implemented to solve that problem mentioned in Task 3.
- A vision of how you are going to apply machine learning to that problem selected in Task 1.

Helpful tip: Complete the following tutorial to develop effective presentation skills: https://roehampton.libwizard.com/f/presentations

- **Part 2: Final Report (75 marks)**
  - Submit a 4000-word final report in PDF format using the provided template.
  - Include your MATLAB (.m/.mlx) file with the report.

5) ASSESSMENT EXPECTATIONS AND RUBRIC:

	Criteria	Expectation	Maximum marks (100)
Presentation (in class) & mid-term report submission	**Mid-term presentation and brief report**	➢ The literature review related to the problem you have opted for from Task 1. ➢ Explaining the dataset by using tools from data visualisation. ➢ Present the descriptive statistical analysis of the dataset you have opted for. ➢ Discuss the part of the solution you have implemented to solve that problem mentioned in Task 3. ➢ Present a vision of how you are going to apply machine learning to that problem opted in Task 1.	25
Final report submission	**Abstract, conclusion and format of demonstration**	A brief 200-300 words glance at 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	A detailed survey of related work that covers relevant literature review on the chosen problem. Discuss the detailed modifications you have conducted to the reference research article.	10
	Mathematical understanding and feature engineering	Detailed mathematical formulation is provided with an appropriate explanation of the algorithmic equations used in the study. Also, showing the 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 are conducted 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 the ability to efficiently visualise overlapping complex data distribution/simulation results in single plots.	5
	Application of machine learning algorithm	Multiple algorithms have been used for comparison purposes, and a comprehensive analysis has been conducted with detailed reasoning.	10
	Programming language used/Statistical software used	A clear well-commented MATLAB code has been developed without using built-in functions.	10

Rubric	**Distribution of marks**				
	100-80%	**79-70%**	**69-60%**	**59-50%**	**49-00% (Fail)**
Abstract, conclusion and format demonstration	A brief 200-300 words glance at the problem statement and its possible solution along with results. Demonstrating report by using appropriate language, clear formatting and correct referencing.	A brief 200-300 words glance at the problem statement and its possible solution along with results. Demonstrating report by using appropriate language, clear formatting and correct referencing.	A brief 200-300 words glance at the problem statement and its possible solution along with results. Demonstrating report by using understandable language, good formatting and correct referencing.	A general 200-300 words glance at the problem statement. Demonstrating report by using understandable language, reasonable formatting and referencing.	The report failed to use understandable language, reasonable formatting and referencing.
Introduction/Literature review appropriateness	Demonstrates outstandingly broad and in-depth independent reading from appropriate sources, including the most current ones in the field. The choice of sources highly enhances the fulfilment of the assignment objectives. Clear, accurate, systematic application of material with highly developed and/or integrated critical appraisal.	Demonstrates very broad and in-depth independent reading from appropriate sources, including the most current ones in the field. The choice of sources clearly enhances the fulfilment of the assignment objectives. Clear, accurate, systematic application of material with well-developed and/or integrated critical appraisal.	Demonstrates in-depth independent reading from appropriate sources, including the most current ones in the field. The choice of sources clearly enhances the fulfilment of the assignment objectives. Clear, accurate, systematic application of material with developed and/or integrated critical appraisal.	Evidence of independent reading from a wide range of appropriate sources, including current ones. Clear, accurate, systematic application of the material. Shows an ability to appraise material critically.	Limited evidence of independent reading. The application of literature is too descriptive overall.
Mathematical understanding and feature engineering	Application of knowledge and understanding of mathematical concepts is outstanding and shows mastery of the discipline and professional practice. Appreciation of the limits of theory is demonstrated throughout the work. The approach to the	Demonstrates a very detailed, accurate, systematic mathematical understanding. Appropriately selected theoretical knowledge is integrated into the overall assessment task, including up-to-date theories, concepts and practices of	Shows a systematic and accurate understanding of key mathematical theories, including the most up-to-date ones, which are appropriately applied, along with own research, within the context of the assessment task.	Effective application of knowledge of key mathematical theories and conclusions resulting from own research.	Knowledge of mathematical theory is inaccurate/incomplete. The choice of mathematical theory is inappropriate/incomplete. Application and/or understanding of concepts are very limited.

	assessment task is informed by the most up-to-date theories, concepts and practices in the discipline and own research.	the subject area and own research.			
Statistical analysis	Appropriate and detailed inferential and descriptive statistical analyses are conducted on the dataset to refine the possible solution.	Appropriate and detailed inferential and descriptive statistical analyses are conducted on the dataset to refine the possible solution.	Appropriate and detailed inferential and descriptive statistical analyses are conducted on the dataset but have limited evidence of refinement of the possible solution	Appropriate inferential and/or descriptive statistical analyses are conducted on the dataset but no evidence is provided of refinement of the possible solution.	Basic descriptive statistical analysis is conducted on the dataset with limited or incorrect or no explanation.
Data visualisation	Various types of graphical representation attempted to visualise the dataset as well as simulation results. Demonstrated the ability to efficiently visualising overlapping complex data distribution/simulation results in single 3D plots.	Various types of graphical representation attempted to visualise the dataset as well as simulation results. Demonstrated the ability to efficiently visualise overlapping complex data distribution/simulation results in single 2D plots.	Various types of graphical representation attempted to visualise the dataset as well as simulation results.	A basic but relevant type of graphical representation attempted to visualise the dataset as well as the simulation results.	The graphical representation is limited and/or inappropriate for the defined problem.
Application of machine learning algorithm	Multiple algorithms have been used for comparison purposes, and a comprehensive analysis has been conducted with detailed reasoning. Demonstrates a good understanding of the application of knowledge in optimising the solution.	Multiple algorithms have been used for comparison purposes, and a comprehensive analysis has been conducted with detailed reasoning. Demonstrates a good understanding of the application of knowledge in optimising the solution.	Multiple algorithms have been used with detailed reasoning. Demonstrates a reasonable understanding of the application of knowledge in optimising the solution.	A single algorithm has been used with limited reasoning. Demonstrates reasonable understanding of the application of knowledge.	A single algorithm has been used with limited reasoning. Demonstrates incorrect or no understanding of the application of knowledge.
Programming language used/Statistical software used	A clear well-commented MATLAB code has been developed without using built-in functions.	A clear well-commented MATLAB code has been developed using minimal built-in functions.	A clear well-commented MATLAB code has been developed using a reasonable number of built-in functions.	A clear well-commented MATLAB code has been developed by using lots of built-in functions.	A vague-commented or uncommented MATLAB code has been developed by using lots of built-in functions.

ADDITIONAL INFORMATION (IF REQUIRED):

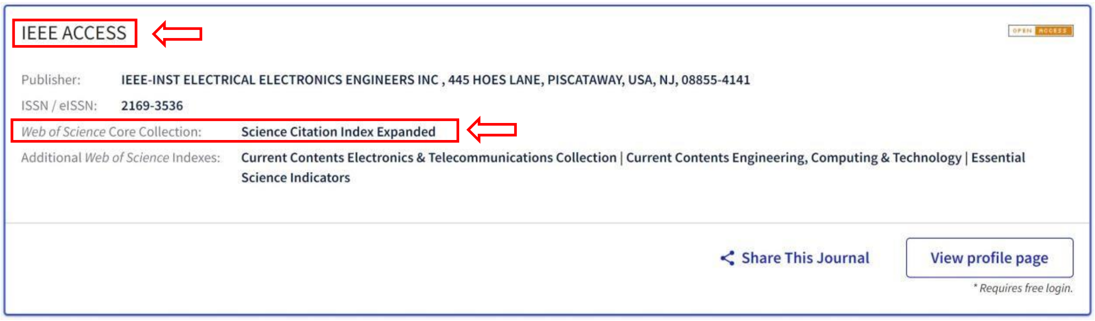
APPNEDIX 1: DOWNLOADING RESTRICTED RESEARCH PAPER:

- For the ScienceDirect paper given in Problem 1, visit: https://library.roehampton.ac.uk/sciencedirect, and for the IEEE paper given in Problems 3, visit: https://library.roehampton.ac.uk/iel
- Use your university credentials to log in, search for the paper title, and download the paper.

APPENDIX 2: HOW TO DECIDE A NEW PROBLEM:

To decide the new problem, you must contact your module tutor first. The reference article for the new problem must belong to the Science Citation Index Expanded (SCIE) database and require images/videos as a dataset. The step-to-step process for deciding a new problem can be conducted by using the following steps:

- Search the keywords of the problem in Google Scholar (https://scholar.google.com/).
- Filter the search result that should not go beyond research paper that are older than 5 years. For example, if you are taking the Application of Data Science module in year 2022, the research article you can opt should lie in the range 2018-2022.
- Discuss the modification you are planning in the reference article with your module tutor.
- Confirm with the tutor, 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.
- Refer to the screenshot below from MJL illustrating the example paper mentioned in Problem 3 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