CST3133 - Coursework: Applied Data Science and AI Workflow

Summary

Version: 1.1

Group project (3 or 4 students per group) - online submission

Deliverables:

- Unified technical report
- Notebok (.ipynb) file that is runnable on Google Colab with the required datasets
    - .ipynb file is a must, Colab link is optional and at student's discretion
- Peer assessment file with mandatory tables

Deadline: 16:00, Wednesday, 16.04.2025 (end of Week 12)

Weighting: 100% of overall mark

Objective

The goal of this coursework is to provide students with hands-on experience in applying the full data science and AI workflow to real-world problems. Students will explore both structured data analysis using machine learning and text-based data processing with natural language processing (NLP) and deep learning techniques. The project emphasizes critical thinking in selecting and preprocessing datasets, designing and evaluating models, and reflecting on ethical considerations such as bias, fairness, and inclusivity.

Dataset

Students are allowed to use any **publicly available** datasets that are suitable for realising the coursework tasks. Students need to upload these datasets and provide the links in relevant sections of the technical report, when submitting the coursework. **Please make sure that the dataset you are planning to use do not require ethical approvals, consents and other permissions**, as

this may significantly delay the coursework submissions. Students are responsible for ensuring compliance with the university's ethical guidelines when selecting datasets for their coursework. Some suggestions on dataset selection:

Truly Public Datasets: If the dataset is explicitly labelled as "open-access," "public domain," or "free to use for any purpose," then no ethical approval is typically needed.

Restricted Access Datasets: If you need to create an account and agree to terms before downloading, the dataset may have usage restrictions (e.g., ***non-commercial use, attribution required, or specific ethical guidelines***).

You may always want to check the ***license or terms of use***:

Creative Commons (CC0, CC-BY, etc.) - Usually okay to use.

Academic Use Only - Cannot be used for commercial projects.

Restricted Use - May require approval from the data provider.

What needs to be submitted

There will be **two submission links**, *one for the unified report in .PDF*, and the other is *a .zip file*. Each student needs to submit:

1. One unified technical report in **.PDF** format including a cover page with student details per group,

and a **separate zip** file that contains:

2. **.IPYNB files** (*it is also a great idea to provide **Colab links** that are accessible in the technical report's References section*) and used **publicly available datasets** (make sure to provide clear links to them in the technical report).
3. **Peer assessment file in word format**. **Each student separately uploads their own peer assessment file**. **This file is not to be shared with the other members of the group.**

Important note: *It is student's responsibility to provide uncorrupted and readable files. Please make sure to check your .zip file and files in it are readable and uncorrupted.* **If the technical document cannot be opened for any reason, it may result in penalties and could be treated as a no submission. This applies to all files, with marking reductions applied**

accordingly. Students are responsible for selecting suitable datasets to complete the tasks effectively, ensuring the work and dataset are not overly simplistic. Most of the codes in the coursework should be runnable on free Google Colab. Students may choose to purchase additional computational resources if desired, but this is entirely optional.

1. Task Breakdown

1.1. Machine Learning Workflow and Ethics (45%)

- **1.1.1 Dataset Selection and Problem Definition (5%)**

  - Select an appropriate structured dataset and clearly define the problem by stating whether it is a classification, regression or clustering problem.
  - Ensure the dataset is relevant and publicly available on GitHub, Kaggle, UCI, etc., and provide a link to it and upload it when submitting the coursework.
  - Provide an overview of the dataset's features and target variable.

- **1.1.2 Data Preprocessing (10%)**

  - Handle missing values appropriately, as required.
  - Scale/normalize numerical features and encode categorical variable, as required.
  - Conduct logical feature selection or engineering, as required.

- **1.1.3 Exploratory Data Analysis (10%)**

  - Create insightful visualizations, e.g., histograms and correlation heatmaps.
  - Identify and comment on trends, patterns, and potential biases in the data.

- **1..1.4 Model Development and Evaluation (15%)**

  - Train and evaluate learning models, e.g., supervised: linear regression, decision trees, and unsupervised: K-means.
  - Use appropriate evaluation metrics, e.g., R-squared, accuracy, precision, recall, F1-score, etc.
  - Provide a clear interpretation of the model performance using classification report, tables or graphs.

- **1.1.5 Ethical Considerations (5%)**

  - Highlight potential biases or fairness issues in the dataset or models.
  - Suggest practical strategies for mitigating these ethical challenges.

1.2. Natural Language Processing and Deep Learning (45%)

- **1.2.1 Text Dataset Selection and Preprocessing (15%)**

  - Select a publicly available text dataset, for example, IMDb reviews, Amazon product reviews, AG News, depending on what you want to achieve.
  - Preprocess the dataset, for example, clean text, tokenize, remove stopwords, etc.
  - Use pre-trained embeddings, e.g., GloVe, Word2Vec for feature representation.

- **1.2.2 Deep Learning Model Implementation (20%)**

  - Design and train a neural network, e.g., RNN, LSTM for a text-based task, e.g., sentiment analysis.
  - Clearly explain the model architecture, e.g., embedding layers, hidden layers, activation functions, and hyperparameter tuning.

- **1.2.3 Evaluation and Insights (10%)**

  - Use evaluation metrics, e.g., accuracy, precision, recall, loss curves.
  - Provide visualizations, e.g., learning curves, confusion matrices, to explain findings, where possible.
  - Highlight strengths, limitations and areas for improvement.

1.3. Report and Code Quality (10%)

- **1.3.1 Report (5%)**

  - Clearly structure the report with task breakdown sections and provide your solutions one-by-one.
  - Include explanations, relevant figures, screenshots and discussions, and proper citations.
  - Adhere to the word limit (max. 7000 words).

- **1.3.2 Code (5%)**

  - Ensure code is well-documented with clear comments.
  - Structure code logically and make it fully runnable on **Google Colab**.
  - Use best practices for coding, including meaningful variable names.
  - **DO NOT include codes** in the unified technical report. You should include cell-output screenshots, as required.

An overview of the assessment requirements for Coursework

Overview of Coursework	
Module code	CST3133
Module title	Advanced Topics in Data Science and Artificial Intelligence
Submission date, time	16:00, Wednesday, 16.04.2025 (end of Week 12)
Feedback type & date	Students receive feedback on the summative assessments within 15 working days of the published deadline. Students also receive formative feedback during the selected laboratory sessions, i.e., ***Milestone 1 in Week 7 and Milestone 2 in Week 12.***
Word count	Max. 7000 words
Assignment type	Unified technical report / group project (highlight of each student's contribution) **"Please note that some students may be invited to attend a VIVA if additional clarification or discussion about their work is needed."**
Assignment structure and format	While writing your report and answering questions, please use the exact same task breakdown order and titles. Numbering of the section can change but the order should NOT. You can directly start with the main structure as highlighted in the coursework task breakdown. We are only interested in the main work and answers provided within the unified technical report. All relevant figures, cell-output screenshots and discussions must be included in the report as it will be the main deliverable to be marked. Do NOT include codes in the report as it will be separately checked in the ipynb file/accessible Colab link. Please use http://www.citethemrightonline.com for anything you would like to cite. Any citation method is fine as long as it is consistently used.
Appropriate use of AI	In this module, you are welcome to use generative AI tools, such as ChatGPT and Copilot to help you write your code and technical report. However, in the technical report, students need to provide: • written acknowledgment of the use of generative artificial intelligence.

	• the extent of use, and how generated materials were used. • descriptions of how the information was generated (including the prompts used). Generative AI can be very useful for brain-storming and many other tasks, but they tend to make mistakes and they may not be able to match the coursework requirements as in the aforementioned task breakdown, which can ultimately lower your marks. Therefore, it is important for students to understand the coursework and its requirements, and do the work on their own. I strongly recommend that students follow the sessions closely and begin their coursework from the very start, working alongside the weekly content. Waiting until the last few days, assuming generative AI can handle everything, often leads to procrastination and unnecessary delays. Starting early will help you stay on track and produce better work on your own.
Assessed learning outcome (s)	LO1, LO2, LO3, LO4, LO5, LO6
Module weighting %	100%
Key reading and learning resources	All reading materials are available at KeyLinks (if the link does not work, please search for the module code: CST3133 on mdx.keylinks.org)

The assessment criteria below illustrate how the 100% scale corresponds to a 20-point scale. Students must document all their findings and answers **in the technical report**, which will be evaluated based on the assessment criteria.

Criteria	**1-4 First**	**5-8 Upper Second**	**9-12 Lower Second**	**13-16 Third**	**17-20 Refer**
	70%+	**60%-69%**	**50%-59%**	**40%-49%**	**Less than 40%**
1.1.1 Dataset Selection and Problem Definition 5%	Excellent problem definition; excellent selection of dataset and features	Clear problem definition with an appropriate dataset; dataset uploaded, link provided and properly cited; well detailed overview of features and target variable.	Relevant dataset and clear problem definition; dataset uploaded, link provided, good explanations of features/target variable.	Dataset is relevant; good problem definition; basic overview provided.	Dataset relevance is limited or unclear; poor problem definition; minimal feature/target details.

1.1.2 Data Preprocessing 10%	Excellent handling of preprocessing, scaling/encoding, justification of feature selection	Missing values handled, features scaled/encoded appropriately, logical feature selection or engineering clearly justified.	Most preprocessing steps completed well; good scaling, encoding, or justification of features.	Preprocessing is done sufficiently; fair scaling, limited justification of features provided.	Preprocessing poorly executed; some issues in scaling, encoding, or missing values.
1.1.3 Exploratory Data Analysis (EDA) 10%	Excellent visualisations; bias identification and explaination	Insightful visualisations with trends, patterns, and biases clearly identified and explained.	Good visualizations with trends/patterns identified; minor issues in clarity.	Basic visualizations provided; limited commentary on trends/patterns.	Minimal visualizations; unclear or incorrect trend/pattern analysis.
1.1.4 Model Development and Evaluation 15%	Excellent handling of model development and evaluations	Models trained and evaluated with appropriate metrics; insightful interpretations with clear visualizations or tables.	Models trained and evaluated; minor issues in metric use or interpretation clarity.	Models implemented but evaluation lacks depth; limited interpretation provided.	Models poorly trained or evaluated; minimal interpretation of results.
1.1.5 Ethical Considerations 5%	Excellent handling of ethical considerations and mitigation strategies	Ethical issues thoughtfully identified and potential mitigation strategies well-discussed.	Ethical issues identified; mitigation strategies present but lack depth or clarity.	Basic ethical issues identified; limited discussion of mitigation strategies.	Minimal ethical discussion; vague or unclear mitigation strategies.
1.2.1 Text Dataset Selection and Preprocessing 15%	Excellent text-dataset selection, preprocessing and embedding application	Relevant text dataset selected; preprocessing (e.g., cleaning, tokenization) done effectively; embeddings applied appropriately.	Text dataset appropriate with basic preprocessing or embedding.	Dataset relevant but preprocessing incomplete or embeddings not fully utilized.	Dataset is not selected or unsuitable or poorly preprocessed; embeddings not utilized.

1.2.2 Deep Learning Model Implementation 20%	Excellent deep learning implementation with all details	Well-designed and trained RNN/LSTM with clear explanation of architecture.	Model implemented and explained; minor issues in training or architecture.	Basic model implementation; limited explanation of architecture.	Poorly implemented model; unclear architecture; poor to no-training.
1.2.3 Evaluation and Insights 10%	Excellent evaluation and findings with amazing visualisations	Relevant metrics used effectively; insightful, nontrivial findings supported by visualizations (e.g., learning curves, confusion matrices).	Metrics and visualizations provided; good interpretation or clarity.	Basic evaluation with fair use of metrics or visualizations; minimal insights provided.	Poor evaluation; unclear metrics; minimal or no insights provided.
1.3.1 Report 5%	Excellent formatting, structuring, cover page with all student names, figures/captions, in-text citations, cross-referencing, to-the-point explanations	Clear structure with task breakdown sections; concise explanations supported by figures, screenshots, and proper citations.	Good report structure with minor issues in clarity, but figures and screenshots are leveraged sufficiently along with good explanations.	Report structured but lacks depth; limited use of figures/screenshots.	Poorly structured report; minimal explanations or figures; exceeds word limit.
1.3.2 Code 5%	Excellent documentation, formatting, logical order, runs well in Colab	Well-documented and logically structured code; runs without issues on Colab; adheres to best practices (e.g., meaningful variable names).	Code mostly functional and clear; minor issues in documentation or structure.	Code somewhat clear but lacks sufficient comments or has minor functional issues.	Poorly documented or disorganized code; significant functional issues.

Mark Allocation Summary

Section	Weighting
Machine Learning Workflow and Ethics	%45
Natural Language Processing and Deep Learning	%45
Report and Code Quality	%10

The following table details the support you will be receiving for this assignment and the feedback opportunities you will have.

Support and draft feedback sessions for Coursework
Coursework briefing We will use online drop-in session in Week 1 for the assessment briefing, where we will explain the assessment requirements and criteria. Coursework will be accessible to students in Week 1 – 23rd January 2025.
Draft feedback opportunities Formative feedback is essential in enhancing students' learning experiences. It provides timely, constructive input on their work while it is still in progress, helping them identify areas for improvement and make necessary adjustments before final submissions. Therefore, we are happy to provide feedback on your progress for the Coursework in selected labs, i.e., ***Milestone 1 in Week 7 and Milestone 2 in Week 12.*** Please make sure to make the most out of these two lab sessions for your progress and to receive formative feedback and be ready to present your progress as it will be very hard to provide any further feedback.

Peer-assessment

Students are required to assess each other's contributions using the following template. Each student needs to assess other members of the group and upload the peer-assessment file. ***A 5% penalty will be applied for failing to upload or submitting an incomplete peer-assessment file, and an additional 5% penalty will be applied if the submitting student does not provide their individual contributions.***

Peer Evaluation Criteria - Rating Scale: 1 (Poor) - 5 (Excellent)**:**

1. **Peer Evaluation of Contributions:** Evaluate your peers on their engagement with all six learning outcomes (LO1-LO6) provided in the module handbook or at the bottom of the peer-assessment template at the coursework page. **This is mandatory.**
2. **Individual Contributions on Learning Outcomes:** Each student must describe their contributions to the project and explicitly address **all six learning outcomes (LO1-LO6)**. **This is mandatory.**
3. **Comment on your Peers:** Provide constructive feedback for each group member. **This is optional.**

Note: A separate word template is provided at the coursework description page, which reveals more details.

Important Guidelines

Coursework will be graded out of 100 points. **It is important to note that individual student contributions will considerably impact the final mark. For example, if the unified report received 100 marks, a student in that group can get as low as 60 marks depending on the averaged peer-assessment points, other penalties, e.g., late submission and no peer-assessment file.** Each student will assess their peers. This peer assessment uses a scale from 1 (poor) to 5 (excellent) for each of the contributions of the student on learning outcomes. **It is recommended that students plan the work at the initial stage with all learning outcomes in mind and work together, communicate findings on each learning outcomes and share/report your findings, select/merge the best findings in the unified technical report**. This will ensure each student covers all the learning outcomes and they get a unified report with all the best part of their findings are merged in the report for maximising the chances of getting highest mark possible. Students is asked to form a group of 3-4 students. **Please ensure that your student numbers, names, and surnames are accurately provided, as they will be required on the cover page of your technical report submission.** ***Only one student from each group must send me an email with all the student details. The deadline for this is 27th of January 11:59 pm.*** **Students who have not formed a group by the deadline will be randomly assigned to a group of four. It is essential that all students actively participate and contribute to their group's work. If a student does not contribute, respond, or communicate with their group members, this must be clearly stated in Section 3 of the peer-assessment file by the other students of the group. In such cases, the non-contributing student may be considered as having not submitted the coursework, and their submission may be deemed invalid or may be invited to a VIVA. Please make every effort to engage with your group members and fulfil your responsibilities to ensure a successful submission.**

Late Submissions

It is critical to submit your work on time. You can submit a coursework up to 24 hours after the deadline without requesting an extension. However, a component grade reduction of equivalent of 10% (or less where this would reduce a pass grade below 40%) will be applied.

Students registered with the Disability and Dyslexia Service, who have a Learning and Support Form, may be able to submit their coursework up to five days late without the mark being capped, if this is indicated as a reasonable adjustment.

If you submit your coursework later than 24 hours after the deadline, your work will not be marked and you will be given a grade of 20 (non-submission) for this piece of assessment. More details can be found at: MDX_HomePage.

Extenuating Circumstances

There may be difficult circumstances in your life that affect your ability to meet an assessment deadline or affect your performance in an assessment. These are known as extenuating circumstances or 'ECs'. Extenuating circumstances are exceptional, seriously adverse and outside of your control. For further information search 'Extenuating circumstances' in MyMDX.

Academic Integrity and Misconduct

Academic Integrity is a set of principles and values to show that you work in a professional, honest and ethical way. You should be aware of the University's academic integrity and misconduct policies and procedures. Taking unfair advantage over other students in assessment is considered a serious offence by the University. Action will be taken against any student who contravenes the regulations through negligence, foolishness or deliberate intent. Academic misconduct takes several forms, in particular:

- **Plagiarism** – using extensive unacknowledged quotations from, or direct copying of, another person's work and presenting it for assessment as if it were your own effort. This includes the use of third-party essay writing services.
- **Collusion** – working together with other students (without the tutor's permission), and presenting similar or identical work for assessment.
- **Infringement of Exam Room Rules** – Communication with another candidate, taking notes to your table in the exam room and/or referring to notes during the examination.
- **Self-Plagiarism** – including any material which is identical or substantially similar to material that has already been submitted by you for another assessment in the University or elsewhere.
- **Unauthorised use of Artificial Intelligence** – using artificial intelligence without referencing as such in your submission. Appropriate use of Artificial Intelligence (AI) is detailed in the assessment requirements grid in section 7.5

Links to the relevant University regulations and additional support resources can be found here: Home Page - MyMDX

Student Success Essentials: This course includes useful information about how to approach your assessments and complete them with honesty. The course also describes what plagiarism (cheating) is and how to avoid it so you don't face any disciplinary action. For successfully completing this course, you will be awarded a certificate that will verify the knowledge you have gained. Certificates can be shared and promoted via LinkedIn and other digital channels. You will have to log into to MyMDX and then MyLearning to access the course. https://mdx.mrooms.net/course/view.php?id=17199

Full details on academic integrity and misconduct and the support available can be found at https://mymdx.mdx.ac.uk/campusm/home#pgitem/419149/t

The Academic Integrity and Misconduct policy is available in our Public Policy Statements (under Academic Quality) at: Our policies | Middlesex University London (mdx.ac.uk)
Referencing & Plagiarism: Suspected of plagiarism?:
http://libguides.mdx.ac.uk/c.php?g=322119&p=2155601
Referencing and avoiding plagiarism:
https://mymdx.mdx.ac.uk/campusm/home#pgitem/419258

The Middlesex University Students' Union (MDXSU) Advice Service offers free and independent support in making an appeal, complaint or responding to any allegations of academic or non-academic misconduct.
https://www.mdxsu.com/advice