National College of Ireland

Masters in Science in Data Analytics

Postgraduate Diploma in Science in Data Analytics

(MSCDAD_A_JAN24I/MSCDAD_B_JAN24I / MSCDAD_C_JAN24I / PGDDA_SEP23)

Data Mining and Machine Learning II

Terminal Assignment-Based Assessment (50%)

Noel Cosgrave, John Kelly, Abubkar Siddig

Release Date: 12th August 2024

Submission Date: 22nd August 2024

Instructions

This assessment is a substitute for the written examination and is designed to evaluate the learning objectives for the module listed below:

LO1 - Critically analyse advanced data mining and knowledge discovery methodologies in order to assess best practice guidance when applied to complex data mining problems

LO2 - Investigate and evaluate key concepts and advanced data mining techniques and assess when to apply such techniques on complex datasets and problem domains.

Answer both of the questions below. Note that each question carries a different number of marks. However both questions will be initially marked out of 100% and then will be weighted according to the number of marks for the question.

Question 1 - Case Study (65 marks)

You have been hired as a data analyst to work on a machine learning project. Based on the last digit of your student number, select the problem domain you will be working in from the table below. You have been asked to propose a methodology suitable for tackling the problem domain. **Make sure you select the correct problem domain**, as marks may not be awarded for incorrect selections.

Last Digit of Student Number	Problem Domain
1 or 2	Defective component detection on printed circuit boards
3 or 4	Identifying anger in audio voice recordings
5 or 6	Question generation in Natural Language Processing
7 or 8	Determining patient outcomes using analysis of EEG signals
0 or 9	Detecting software vulnerabilities in source code

You should then prepare a report, 5 to 6 pages in length and using Harvard Referencing that presents your assumptions, proposed approach and rationale with respect to the aspects of the project listed in the rubric on the following page.

Tips

- Remember that you are proposing a methodology. You do not need to implement any aspect of the project.
- In order to properly answer this question, you will need to carry out a literature review covering (only) the recent work carried out in the domain. Although this should inform your approach, your proposed methodology should not merely be a carbon-copy of that proposed by a cited (or uncited) author.
- If a particular aspect listed above is not applicable, or is implicit in the machine learning approach taken, then ensure you state this, citing evidence from the literature if necesssary.
- Make sure you state your assumptions. Assumptions must be plausible, i.e. the assumption that there are no missing or out of range data will lose you marks if such data are a known problem in the domain.
- If a particular work in the domain has informed your choice of approach, it must be appropriately cited, with any text from the cited work properly summarised or paraphrased.
- Avoid any lengthy discussion on how any particular technique works. Instead you should present a rationale for why the technique is the most appropriate for the problem at hand.
- The report should be entirely in your own words. Please avoid using thesauri, as synonyms do not always carry precisely the same meaning and can often render a sentence unintelligible.
- At each step in the methodology you may have several appropriate choices. You should indicate not only why you selected one over the others, but also (briefly) why you rejected those that you did not use.
- If there are particular challenges that you expect will be encountered, make sure you state them, citing relevant literature where appropriate.

GRADING RUBRIC - Question 1						
CRITERION	*Upper H1*	*H1*	*H2.1*	*H2.2*	*PASS*	*FAIL*
Exploratory Data Analysis / Data Preparation 15%	The approach to EDA and data preparation is highly appropriate and exceptionally well justified using suitable literature or underlying theory.	The approach to EDA and data preparation is appropriate and very well justified using suitable literature or underlying theory.	The approach to EDA and data preparation is appropriate and well justified using suitable literature or the underlying theory.	The approach to EDA and data preparation is largely appropriate and reasonably well justified sing suitable literature or underlying theory.	The approach to EDA and data preparation is somewhat appropriate and adequately justified but with only rudimentary use of suitable literature or underlying theory.	The approach to EDA and data preparation is missing or inappropriate and/or poorly justified.
Dimensionality Reduction / Feature Selection 15%	The need for (or lack of same) for dimensionality reduction and/or feature selection is correctly identified and exceptionally well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for dimensionality reduction and/or feature selection is correctly identified and very well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for dimensionality reduction and/or feature selection is correctly identified and well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for dimensionality reduction and/or feature selection is largely correctly identified and reasonably well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for dimensionality reduction and/or feature selection is somewhat identified and adequately discussed but with only rudimentary use of suitable literature or underlying theory.	The discussion on dimensionality reduction is missing and/or poorly justified. No use of suitable literature or underlying theory.
Feature Engineering / Feature Extraction 10%	The need for (or lack of same) for feature engineering and/or feature extraction is correctly identified and exceptionally well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for feature engineering and/or feature extraction is correctly identified and very well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for feature engineering and/or feature extraction is correctly identified and well discussed using suitable literature or the underlying theory.	The need for (or lack of same) for feature engineering and/or feature extraction is largely correctly identified and reasonably well discussed but with only rudimentary use of suitable literature or underlying theory.	The need for (or lack of same) for feature engineering and/or feature extraction is somewhat identified and adequately discussed but with only rudimentary use of suitable literature or underlying theory.	The discussion on feature engineering and/or feature extraction is missing or mostly incorrect. No use of suitable literature or underlying theory.
Choice of modelling techniques 20%	The choice of modelling techniques is exceptionally appropriate and is exceptionally well discussed using suitable literature or the underlying theory.	The choice of modelling techniques is very appropriate and is very well discussed using suitable literature or the underlying theory.	The choice of modelling techniques is appropriate and is well discussed using suitable literature or the underlying theory.	The choice of modelling techniques is appropriate and is reasonably well discussed but with only rudimentary use of suitable literature or underlying theory.	The choice of modelling techniques is somewhat appropriate and is adequately discussed but with only rudimentary use of suitable literature or underlying theory.	The discussion on the choice of modelling technique is missing or mostly incorrect. No use of suitable literature or underlying theory.
Hyperparameter Optimisation 15%	The suggested hyperparameter optimisation technique is exceptionally appropriate and is exceptionally well discussed using suitable literature or the underlying theory.	The suggested hyperparameter optimisations technique is very appropriate and is very well discussed using suitable literature or the underlying theory.	The suggested hyperparameter optimisation technique is appropriate and is well discussed using suitable literature or the underlying theory.	The suggested hyperparameter optimisation technique is appropriate and is reasonably well discussed but with only rudimentary use of suitable literature or underlying theory.	The suggested hyperparameter optimisation technique is somewhat appropriate and is adequately discussed but with only rudimentary use of suitable literature or underlying theory.	The discussion on hyperparameter optimisation is missing or mostly incorrect. No use of suitable literature or underlying theory.
Model Evaluation 10%	The selected evaluation approach is exceptionally appropriate and is exceptionally well discussed using suitable literature or the underlying theory.	The selected evaluation approach is very appropriate and is very well discussed using suitable literature or the underlying theory.	The selected evaluation approach is appropriate and is well discussed using suitable literature or the underlying theory.	The selected evaluation approach is appropriate and is reasonably well discussed but with only rudimentary use of suitable literature or underlying theory.	The selected evaluation approach is somewhat appropriate and is adequately discussed but with only rudimentary use of suitable literature or underlying theory.	The discussion on model evaluation is missing or mostly incorrect. No use of suitable literature or underlying theory.
Scalability Issues 7.5%	Scalability issues are fully identified and exceptionally well discussed.	Scalability issues are fully identified and very well discussed.	Scalability issues are largely identified and well discussed.	Scalability issues are largely identified and reasonably well discussed.	Scalability issues are partially identified and adequately discussed.	The discussion on scalability issues is missing or mostly incorrect. No use of suitable literature or underlying theory.
Ethical Implications 7.5%	Ethical implications are fully identified and exceptionally well discussed.	Ethical implications are fully identified and very well discussed.	Ethical implications are largely identified and well discussed.	Ethical implications are largely identified and reasonably well discussed.	Ethical implications are partially identified and adequately discussed.	The discussion on scalability issues is missing or mostly incorrect. No use of suitable literature or underlying theory.

Question 2 - Paper Review (35 marks)

Based on the **last digit** of your student number, you are required to review the respective paper from the table below. All papers are available on the Moodle page. **Please make sure you review the correct paper, as marks may not be awarded for reviews of the wrong paper**.

As part of the review you must complete an evaluation report, using the evaluation criteria outlined in the Elsevier reviewer guidelines (supplied on Moodle). Note that you are not required to verify any statistical results.

Your review should be approximately 2½ to 3 pages in length.

Last Digit of Student Number	Paper To Review
0 or 1	Efficient Multi-hop Question Generation
2 or 3	A Novel Deep Multi-head Attentive Vulnerable Line Detector
4 or 5	Violence Detection in Real-Life Audio Signals Using Lightweight Deep Neural Networks
6 or 7	Towards the Prognosis of Patients in Coma using Echo State Networks for EEG Analysis
8 or 9	Multi-Scale Vision Transformer for Defect Object Detection

Tips

- You do not need to cite the paper you are reviewing.
- In addition to stating your conclusions, remember to say how you reached those conclusions, i.e. present evidence from the paper you are reviewing that led you to your decision.
- The review should be in your own words. Please avoid using thesauri, as synonyms do not always carry the same precise meaning and can render a sentence unintelligible.
- If a particular aspect of the paper is either missing entirely or present but suboptimal, offer your opinion on how it could be improved.
- Note the marks awarded for each section. This is an indication of the amount of work and level of detail required.

Marking

Your review of the paper will be marked according to the rubric on the following page:

GRADING RUBRIC - Question 2						
CRITERION	*Upper H1*	*H1*	*H2.1*	*H2.2*	*PASS*	*FAIL*
Structure and title 2%	The appropriateness of the structure and suitability of the title have been exceptionally well identified and fully discussed.	The appropriateness of the structure and suitability of the title have been very well identified and fully discussed.	The appropriateness of the structure and suitability of the title have been well identified and discussed.	The appropriateness of the structure and suitability of the title have been reasonably well identified and discussed.	The appropriateness of the structure and suitability of the title have been adequately identified and discussed.	The appropriateness of the structure and suitability of the title have been poorly identified and discussed.
Abstract 3%	The quality of the abstract has been exceptionally well discussed.	The quality of the abstract has been very well discussed.	The quality of the abstract has been well discussed.	The quality of the abstract has been reasonably well discussed.	The quality of the abstract has been adequately discussed.	Poor or no discussion on the quality of the abstract.
Introduction 5%	The quality of the introduction has been exceptionally well discussed.	The quality of the introduction has been very well discussed.	The quality of the introduction has been well discussed.	The quality of the introduction has been reasonably well discussed.	The quality of the introduction has been adequately discussed.	Poor or no discussion on the quality of the introduction.
Graphical abstracts and/or highlights 5%	The use and appropriateness of graphical abstracts have been exceptionally well discussed.	The use and appropriateness of graphical abstracts have been very well discussed.	The use and appropriateness of graphical abstracts have been well discussed.	The use and appropriateness of graphical abstracts have been reasonably well discussed.	The use and appropriateness of graphical abstracts have been adequately discussed.	Poor or no discussion on graphical abstracts and their use.
Methodology 25%	The suitability, robustness and reproducibility of the methodology have been exceptionally well discussed.	The suitability, robustness and reproducibility of the methodology have been very well discussed.	The suitability, robustness and reproducibility of the methodology have been well discussed.	The suitability, robustness and reproducibility of the methodology have been reasonably well discussed.	The suitability, robustness and reproducibility of the methodology have been adequately discussed.	Poor or no discussion on the suitability, robustness and reproducibility of the methodology.
Results 25%	The clarity and interpretation of the results have been exceptionally well discussed.	The clarity and interpretation of the results have been very well discussed.	The clarity and interpretation of the results have been well discussed.	The clarity and interpretation of the results have been reasonably well discussed.	The clarity and interpretation of the results have been adequately discussed.	Poor or no discussion on the clarity and interpretation of the results.
Conclusion/Discussion 25%	The plausibility of the conclusions and quality of the discussion have been exceptionally well discussed.	The plausibility of the conclusions and quality of the discussion have been very well discussed.	The plausibility of the conclusions and quality of the discussion have been well discussed.	The plausibility of the conclusions and quality of the discussion have been reasonably well discussed.	The plausibility of the conclusions and quality of the discussion have been adequately discussed.	Poor or no discussion on the plausibility of the conclusions and quality of the discussion .
Language 5%	Language errors (if any) have been completely identified and thoroughly discussed.	Language errors (if any) have been completely identified and well discussed.	Language errors (if any) have been largely identified and reasonably well discussed.	Language errors (if any) have been partially identified and reasonably well discussed.	Language errors (if any) have been partially identified and adequately discussed.	Language errors (if any) have not been identified or discussed.
Previous Research 5%	Ethical implications are fully identified and exceptionally well discussed.	Ethical implications are fully identified and very well discussed.	Ethical implications are largely identified and well discussed.	Ethical implications are largely identified and reasonably well discussed.	Ethical implications are partially identified and adequately discussed.	The discussion on scalability issues is missing or mostly incorrect. No use of suitable literature or underlying theory.

Submission

Both the report and review should be uploaded as a **single document** in **PDF format only** to the Turnitin link on Moodle by the submission date shown at the top of this document. As this is a terminal assessment, late submissions will not be accepted.

Academic Integrity

This is an **individual assessment** and an exam replacement. As such your submission must be entirely your own work. Collaboration with others, whether fellow students or not, is **strictly prohibited** under any and all circumstances. This means that not only should answers to questions not be shared with others, but approaches to tackling questions should also not be discussed. In the event that you require clarification on any aspect of this assessment, please contact your lecturer, who will be happy to help clear up any doubts.

Any written work created by others must be properly cited and should be paraphrased or summarised where possible, otherwise it should be included in quotes. Figures not created by you should include an acknowledgement detailing the name(s) of the creator(s).

Although AI tools may be used to help you find suitable papers, any other use of AI tools such as ChatGPT, Quillbot or similar is **strictly prohibited**.

Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library [1].

Note: All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation. Any submission showing evidence of such misconduct will be referred to the college's academic misconduct committee for disciplinary action.

[1] https://libguides.ncirl.ie/workingremotely/academicintegrity