National College of Ireland

H9DAI: Data Analytics for Artificial Intelligence

MSc in Artificial Intelligence, 2024-25, MSCAI_JAN25
Postgraduate Diploma in Artificial Intelligence, PGDAI_SEP24

Submission Deadline: Friday 14th April 2025

Continuous Assessment (CA) Type: Project 70%

Dr Anu Sahni / Dr Rejwanul Haque

Weight: This Terminal Assignment-Based Assessment (TABA) is worth 70% of the final grade awarded.

Instructions: The project is equivalent to the terminal exam and is worth 70% of the total assessment. Students will work in groups of up to 4 on a problem. Each member will have to carry out all the tasks from selection of data through pre-processing to implementation of a machine learning model to evaluation and presentation of the results on his/her dataset. The datasets from different members of a group can be related or complementary. The project meets all the Learning Outcomes of the module:

LO1 Demonstrate critical understanding of data analytics and machine learning concepts and methods.

LO2 Employ data analytics tools to manipulate, synthesise, explore, and visualise data.

LO3 Select and apply machine learning concepts and methods to assist on business decision-making.

LO4 Evaluate and employ graphical tools for building comprehensive analytics processes and dashboards.

LO5 Critically analyse, compare, summarise, and present results to support decision making and address requirements in real-world business problems

Learners will have to identify and carry out a series of analytic tasks on up to 4 large datasets depending on the number of members in the group. The dataset can be a collection of datasets that are somehow related or complement each other. You will utilise appropriate tools and techniques for data extraction, processing, analysis and critical evaluation. The final submission will consist of an academic research paper style report as well as the implemented data analytics artefact. It is also expected students to present and communicate the results/insights of their study.

The over-arching focus of the project is to develop a portfolio of methods that can reveal insights into the performance and application limitations of machine learning methods in different contexts. The application of each method should be applied in order to answer a specific (small-scale) research question aligned to the overall goal(s) of the project. It is also expected that the application of each method is accompanied by an appropriately sized lit review documenting pertinent and contemporary approaches in the literature that can both inform the application of a method as well as justify its potential merit(s).

Projects will be assessed based on their novelty, technical quality, potential impact, insightfulness, depth, clarity, and reproducibility. Code and data sets are to also be submitted with the paper. Algorithms and resources used in a paper should be described as completely as possible to allow reproducibility. This includes experimental methodology, empirical evaluations, and results. The reproducibility factor will play an important role in the assessment of each submission.

SUBMISSION DETAILS:

- Deadline for submission
  : Friday 14th April 2025 23:59
-

TURNITIN: All report submissions will be electronically screened for evidence of academic misconduct (i.e., plagiarism and collusion)

Key details, requirements, and definitions

Data Requirements: each group should work upon 2 to 3 large datasets depending on the number of members in the group. The dataset can be a collection of datasets that are somehow related or complement each other dataset. Each dataset should be for predictive analytics tasks, i.e. it should have a meaningful easily identifiable response variable. Each dataset should also be suitably large (at least 10000 rows, and at least 10 columns). An example dataset meeting these requirements is the Adult dataset available here: https://archive.ics.uci.edu/ml/datasets/adult

Number of methods : in total, you should apply and critically evaluate at least 3 to 4 methods of machine or statistical learning for this project to facilitate your discussion.

Notions of performance : the discussion of performance should be orientated around multiple notions of performance. It is not sufficient to discuss only accuracy or R2 for the methods applied. Other possibilities include, but are not limited to: Cohen's Kappa, RSME, RSS, Sensitivity/Specificity, F-Measure, and MAPE.

Methodology : the application of each method must follow an appropriate data mining methodology, where CRISP-DM [1] and KDD [2] are foreseen as most likely to be appropriate.

It is essential that projects unambiguously evidence all of the following.

1. A critical analysis of fundamental data analysis concepts to address requirements in real-world business problems
2. The extraction, transformation, exploration, and cleaning of datasets in preparation for the datamining and machine learning methods used in the project.
3. The building and evaluation of machine learning models on a dataset.
4. The visual representation the findings, interpretation and evaluation of information and knowledge that is drawn from the datasets as a central theme in the project.
5. Critically review, compare, summarise, and present results to support decision making and
6. References

Final Report

The final report must follow the IEEE conference format and should be up 8-10 double column pages in length (this includes all figures and references). For this exercise IEEE style referencing, not Harvard referencing, should be used. Papers over 10 pages may be subjected to a 5 percentile point penalty, i.e. the maximum mark for the paper will be 95%. Word and LATEX templates are available here:
http://www.ieee.org/conferences_events/conferences/publishing/templates.html
Your report should discuss your approach with respect to the application of CRISP-DM [1] or KDD [2], with an emphasis on the critical evaluation of the methods selected. The following structure is suggested for the report (see Table 1 for more detail):

1. **Abstract and Objectives (10)**: 150-250 words providing a high-level of the project, its core findings, and the domain of the datasets (not necessarily in this order). The remainder of 1st page (+ up to 1 column) should present and discuss the problem / research question(s) / objective(s) of the project and (optionally) provide a concise overview of the following sections (max 1-2 lines per each).

2. **Related Work (10)**: 1 or 2 pages (between 12 - 20 references in total) – this should not only summarise related work, but also critically evaluate (positive and negative aspects) of key related work with respect to the topic and domain of the project, i.e. how well/badly does the related work artefact address your question(s) / objective(s), what aspects are useful to consider, what are the limitations etc. Also include here a discussion on the previous uses of the datasets and the methods applied. If you plan to reuse a method already applied to this dataset, discuss what you expect to gain by doing this. If you are unsure about how to write a literature review, or generally would like to see what one looks like, see [3].

3. **Dataset and Choice of Methods (15)**

   Data Requirements Data Requirements: Each dataset should be for predictive analytics tasks, i.e. it should have a meaningful easily identifiable response variable. Each dataset should also be suitably large (at least 10000 rows, and at least 10 columns).

   Number of methods : in total, you should apply and critically evaluate at least 3 to 4 methods (2 to 3 members group), of machine or statistical learning for this project to analyse and make the prediction of the problem.

4. **Methodology and Pre-processing (20)**: The description of dataset and how have you approached answering your question. Additional (technical) details can also be discussed here. Essentially, you should recount how you applied either CRISP-DM [1] or KDD [2] (but not both) to facilitate your research question(s). You should also include here a discussion on key preliminary aspects of the methodology, such as how the datasets have been prepared for study (i.e., the pre-processing, and transformation stages). This section should include the extraction, transformation, exploration, and cleaning of datasets in preparation for the datamining and machine learning methods used in the project.
5. **Evaluation and Presentation (30)**– what performance measures have you selected and why (discuss how the choice of performance measures is appropriate You should also present, visualising, and discuss the results in detail in this section: what are their implications? What do they show / not show? Etc. A discussion on sampling methods is expected here too.

6. **Conclusions and future work (10)**: summarise your findings, and discuss limitations / extensions that were you to have more time, you would do next to improve / extend your study. Summarise the (partial) answer to the research question(s) at a high level, and note the key implications of your findings with respect the methods studied.

7. **References and Presentation (5):** Include a list of references used in your report. Note that websites are not references, they should be referred to in footnotes. All referenced works should be locatable in Scopus. Do not use papers from any of the sources noted in this list: https://beallslist.weebly.com; these papers

may be plagiarised, low in quality, not subject to rigorous (or any appropriate) peer review, and should generally be held as dubious and untrustworthy. Note that typically, if a paper is in Scopus, it is unlikely to be in this list.

Presentation

Presentations will be conducted during class, with the following mandatory requirements:

Max length : 5 mins

Overview : give a quick overview of your dataset and research questions

Demo : Recreate (run the code) and discuss the most significant results of the project

Potential Sources of Data

Possible sources of datasets include, but are not limited to:

- Statista https://www.statista.com
- European Data Portal, EU Open Data Portal, and other http://data.europa.eu/
- UK's open government data repository: http://data.gov.uk
- Central Statistics Office, Ireland: http://www.cso.ie
- Kaggle: http://www.kaggle.com
- Run My Code: http://www.runmycode.org/
- Amazon's public dataset repository: https://aws.amazon.com/datasets
- Google's Public Data Directory: http://www.google.com/publicdata/directory
- The UCI machine learning repository: http://archive.ics.uci.edu/ml/
- Google Data Search: https://toolbox.google.com/datasetsearch
- Zenodo https://zenodo.org
- Dublinked https://data.smartdublin.ie
- Data.gov https://www.data.gov/
- Quandl https://www.quandl.com

References

[1] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. 2000.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, 1996.

[3] M. Hall, A. Mazarakis, M. Chorley, and S. Caton. Editorial of the special issue on following user pathways: Key contributions and future directions in cross-platform social media research. *International Journal of Human Computer Interaction*, 2018.

Grading Rubric

Table 1: Grading Rubric

Criteria	High H1	H1	H2-1	H2-2	Pass	Fail
1. Abstract and Objectives (10%)	Very challenging project objectives are well presented, met and thoroughly motivated as well as discussed.	Challenging project objectives are well presented, met and thoroughly motivated as well as discussed.	Appropriate project objectives are well presented, met and thoroughly motivated as well as discussed.	Appropriate project objectives are presented, mostly met and motivated as well as discussed.	There are clear objectives, which are at least partially met.	Cannot discern project objectives, and/or if project objectives were met.
2. Related work (10%)	Discussion of related work is excellent, and the choice of papers to discuss excellently situates the project within the literature.	Discussion of related work is v. good, and the choice of papers to discuss excellently situates the project within the literature	Discussion of related work is good and the choice of papers to discuss well situates the project within the literature	Discussion of related work is appropriate and the choice of papers to discuss well situates the project within the literature	Discussion of related work is appropriate, and the choice of papers appropriately situates the project within the literature	Discussion of related work lacks depth, and/or the choice of papers seems somewhat arbitrary.

DRAFT

3. Dataset and Choice of methods (15%)	The student has studied a selection of complex methods illustrating a well thought out approach to addressing their objective(s).	The student has studied some complex methods illustrating a well thought out approach to addressing their objective(s).	At least two methods requires the application of advanced methods.	At least one methods requires the application of advanced methods.	The student has appropriately selected methods to address their objective(s), but played it safe.	Choice of methods appears arbitrary, or not well justified.
4. Methodology and Pre-processing (20%)	It is hard to find fault in the approach.	All stages of KDD/CRISP-DM are rigorously applied.	All stages of KDD/CRISP-DM are rigorously applied. Some minor shortcuts or errors may be present.	All stages of KDD/CRISP-DM are appropriately applied, but the general approach lacks some depth. There may be some mistakes in the approach taken.	All stages of KDD/CRISP-DM are appropriately applied, but the general approach lacks depth. There may be significant mistakes in the approach taken.	KDD or CRISP-DM not appropriately followed and/or applied. The approach taken may also be hard to discern.
5. Evaluation and Presentation (30%)	All key decisions are justified with the state-of-the-art visualisation.	All key decisions are justified with outstanding visualisation.	Most key decisions are justified with excellent visualisation.	Key decisions are justified with good visualisation., but more depth is needed.	Some key decisions are justified with with some visualisation., but more depth is needed.	Key decisions are not justified or substantiated with inappropriate literature.
6. Conclusions and Future Work (10%)	Insightful conclusions, which appreciate key limitations and implications of the project. Key implications of the	Insightful conclusions, which appreciate limitations and implications of the project. Implications of the project are	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Appropriate future	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Future work lacks	Implications and limitations not well understood. Future work lacks depth and creativity, but is appropriate.	Implications and limitations not understood. Future work seems arbitrary or inconsistent with project findings

DRAFT

	project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	anchored with relevant literature. Well-conceived and thought out future work is discussed.	work is discussed and presented.	depth and creativity, but is appropriate.		
7. References and Presentation (5%)	Exceptionally well written, and presented, with no mistakes in formatting or referencing.	Well written, with no (large) language errors. All figures are well conceived and readable. The IEEE template is adhered to. Report does not exceed the length limits. References are appropriately and correctly used.	Main document has a few language and/or style errors. Figures are well presented. IEEE template and length limit are adhered to. References are complete, and correctly used.	Main document is readable with some language and/or style errors. Some figures are mostly well presented. IEEE template is largely adhered to. References are mostly complete and correctly used.	Main document is readable with some language and/or style errors. Some figures may be hard to read or presented in a suboptimal manner. IEEE template is largely adhered to. References are mostly complete and correctly used.	Littered with typos, and/or poor use of English. IEEE template may have been broken. Figures may be hard to read. References (if any) are probably incomplete.

DRAFT

DRAFT