

National College of Ireland

**(MSCDAD\_A/** **MSCDAD\_B/** **MSCDAD\_C)**

**Release Date: 14/10/24**

**Due Date: 13/12/24**

**M. Bradford/A. Sahni/S. Zahoor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Data Mining & Machine Learning**

**Assessment Type: Project**

**Weight:** The assignment will be marked out of 100. This assessment contributes towards a maximum of 100% of the module’s marks.

**Instructions:**

* Upload your project report, code, and any other relevant artefacts to Moodle via the relevant submission links.
* **This is an individual assessment – the project MUST be your own work.**
* **Include assessment submission cover pages as part of your submission document.**
* **Please familiarize yourself with the college’s policies on academic honesty and the use of AI.**

**SUBMISSION DETAILS**:

Submit the project report and all file(s)/notebooks/presentation links to the Moodle Project Submission links before the deadline.

**TURNITIN**: All report submissions will be electronically screened for evidence of academic misconduct (i.e., plagiarism and collusion). Please include assessment submission cover pages.

**Duration of Continuous Assessment:** Project

**[If Applicable] Attachments:**

# Introduction

This project is designed to evaluate the learning objectives of the Data Mining & Machine Learning module.

**Project Overview**

Produce a portfolio of studies that critically compare the performance of different machine learning methods applied to at least 3 large datasets. One of these datasets must predominantly contain text data.

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.

**Key details, requirements, and definitions**

**Due:** 13th December 2024.

**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 (i. at least 10000 rows and at least 10 columns for datasets with predominantly numeric data; ii) at least 10000 rows of data for the predominantly text-based dataset with the text data containing at least 40 words per row).

**Deliverables:** There are 3 deliverables for this project:

1. a pdf final report;
2. a .zip containing all code, and the datasets used;
3. a 7 min video presentation illustrating key parts of the project.

**Number of methods**: in total, you should apply and critically evaluate at least 5 methods of machine or statistical learning for this project to facilitate your discussion. At least one predictive text analytics/machine learning scheme must be applied to your text-based dataset.

**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, FMeasure, 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 mining and knowledge discovery methodologies in order to assess best practice guidance when applied to data mining problems in the specific context of the project.
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 data mining and machine learning models on a variety of datasets.
4. The extraction, interpretation and evaluation of information and knowledge that is drawn from the datasets as a central theme in the project.
5. The critical review of relevant data mining research to afford the assessment of research methods applied in the project.

# Final Report [100%]

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 penalty. 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 Grading Rubric for more detail):

**Abstract:** 150-250 words providing a high-level of the project, its core findings, and the domain of the datasets (not necessarily in this order).

**Introduction:** remainder of 1st page (+ up to 1 column). Should motivate the work, present and discuss the research question(s) / objective(s) of the project and (optionally) provide a concise overview of the following sections (max 1-2 lines per each).

**Related Work:** 1 or 2 pages (15 or more 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].

**Data Mining Methodology (can be named differently):** 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).

**Evaluation:** How have you used your method(ology) to answer the question (evaluation methodology), i.e. how do you know that a method is good? I.e. what performance measures have you selected and why (discuss how the choice of performance measures is appropriate). If you have to parametrise part of an approach, how have you done that, and why were these choices made, and what impacts can different parameterisations have on your results? You should also 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.

**Conclusions and future work:** 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.

**References:** 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.

# Video Presentation [Mandatory Submission]

Presentations will be conducted via a video presentation, with the following mandatory requirements:

**Max length:** 7 mins

**Methodology:** give a quick overview of your methodology

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

**Upload:** The video should be submitted as an mp4 file or a link to the video must be privately shared (included in the .zip archive of code)

**Materials:** Any used in the presentation should be uploaded to Moodle (included in the .zip archive of code).

# 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.

# Marking

The project will be marked according to the grading rubric provided at the end of this document.

# 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 all circumstances. 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). Code found on the internet should not be claimed as your own, but instead a comment should be included in the source code indicating where you obtained it. The use of large language models such as ChatGPT or AI coding assistants for any part of the assessment is strictly prohibited. Students are strongly advised to familiarise themselves with the Guide to Academic Integrity produced by the NCI Library and the college’s policy on the use of AI.

**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.

**Grading Rubric**

**A white paper with black text

Description automatically generated**

**A white paper with black text

Description automatically generated**

**A close-up of a text

Description automatically generated**

**A close-up of a white background

Description automatically generated**

**A white text on a white sheet

Description automatically generated with medium confidence**

**A close-up of a text

Description automatically generated**