Project Description 2019–2020
Data Mining and Machine Learning I

Luis Gustavo Nardin

4th February 2020

1 Administrative Data

This project is an **individual Project assignment** that is worth **100% of the final grade** awarded. As published in the School portal, the deadline for submission of this assignment is divided into two parts:

Part I: Thursday 12^{th} March 2020 23:59hrs
Part II: Friday 24^{th} April 2020 23:59hrs

Extension/Re-run

Should any student miss the assessment deadline with a valid reason, he/she can now apply for an application for coursework Extension/Re-run Form online, via NCI360.

N.B.

All submissions will be electronically screened for evidence of academic misconduct (plagiarism and collusion).
This coursework represents 100% of your overall repeat mark for this module.

2 Project Overview

Produce a portfolio of studies that critically compare the performance of different machine learning methods applied to at least 3 large datasets.

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 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 must also be submitted with the report. Algorithms and resources used in a report 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.

3 Key details, requirements, and definitions

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 10,000 rows**, and **at least 10 columns**). An example dataset meeting these requirements is the Adult dataset available at `https://archive.ics.uci.edu/ml/datasets/adult`

Deliverables: There are **4 deliverables** for this project:

1. **Proposal/Interim Progress Report** (PDF format)
2. **Final Report** (PDF format)
3. **Source-code and datasets** used in the project as a compressed ZIP file
4. **Video Presentation** of 7 minutes in length illustrating the key parts of the project.

NOTE 1: Ensure that your name in full (as per NCI official documents) and student number are clearly visible on the front page.

NOTE 2: Name your files starting with the first letter of your given name followed by the first three letters of your surname, your student id, a dash, and the text 'report-dmml1'. No spaces should be included in the filename. That is to say, when "Mary Murphy" with student ID 15123456 submits her assignment, she should name the file `mmur15123456-report-dmml1.pdf`.

Effort: It is expected that this project require approximately between 50–80 hours of work.

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.

Notions of Performance: The discussion of performance should be orientated around multiple notions of performance. It is not sufficient to discuss only accuracy or R^2 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 the project shows unambiguously evidence of:

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 data mining 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.

4 Proposal / Interim Progress Report [20%]

The proposal should pitch your project idea and provide an overview of the methods intended to be used to develop and evaluate the chosen machine learning methods. The report should be structured according to the sections in Table 1.

Table 1: Proposal / Interim Progress Report sections structure.

Section	Weight	Length Limit	Description
Motivation	25%	Max 300 words.	Motivate your choice of topic.
Research Question	15%	Max 50 words.	What question do you plan to answer?
Initial Review	25%	Max 400 words.	Brief review of pertinent literature. Max of 12 papers.
Data Sources	10%	List of data sources.	Source of data and brief description.
Machine Learning Methods	10%	At least one per data source.	Brief description of each method and rationale for choosing the method.
Evaluation Methods	10%	Max 100 words per method.	Brief description of proposed methods to evaluate chosen machine learning method.
Bibliography	5%	Exempt from length limits.	

5 Final Report [80%]

The final report must follow the IEEE conference format and should be **between 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 will be subjected to a 5% point penalty, i.e., the maximum mark for the paper will be 95%. Microsoft Word and LaTeX templates are available at `http://www.ieee.org/conferences_events/conferences/publishing/templates.html`.

The final report should include a discussion of the 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 final report:

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

Introduction: Remainder of 1^{st} page (+ up to 1 column in the 2^{nd} page). It 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 (20 or more references in total) – this should not only summarize the related works, but also critically evaluate their key positive and negative aspects with respect to the topic and domain of the project, i.e., how well/badly do the related works 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: This section can be named differently. But it should describe 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? what performance measures have you selected and why (discuss how the choice of performance measures is appropriate). If you have to parametrize part of an approach how have you done that, and why were these choices made, and what impacts can different parameterizations 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: Summarize your findings, and discuss limitations / extensions that were you to have more time, you would do next to improve / extend your study. Summarize 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 plagiarized, 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.

Although not recommended, you may remove, change and/or alter methods, datasets and/or key research question(s) or objective(s) after submission of the Proposal / Interim Progress Report.

6 Video Presentation

Elaborate a video presentation with the following mandatory requirements:

Max length: 7 minutes

Methodology: Briefly present the methodology used to develop the project

Demo: Recreate (i.e., run the code) and discuss the most significant results of the project

Failed Methods: Demonstrate one approach that was excluded from your final report and discuss why

Upload: The video and any material used to prepare the presentation have to be uploaded on Moodle

7 Potential Data Sources

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

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

8 Marking Grid

Worth 100% of the final mark, this coursework will be graded using the marking grid shown in Table 2.

Table 2: Marking Grid – Project, Data Mining and Machine Learning I 2019–2020

CRITERIA	HIGH H1	H1	H2.1	H2.2	Pass	Fail
Objectives and Motivation. (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.
Discussion and 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 very 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, or the choice of papers seems somewhat arbitrary.
Choice of Methods. (10%)	The student has studied a selection of complex methods illustrating a well thought out approach to addressing their objective(s).	The student has studied a selection some complex methods illustrating a well thought out approach to addressing their objective(s).	Application of at least two advanced methods.	Application of at least one advanced method.	The student has appropriately selected methods to address their objective(s), but played it safe.	Choice of methods appears arbitrary, or not well justified.
Methodology. (25%)	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 or applied. The approach taken may also be hard to discern.
Evaluation. (25%)	All key decisions are justified with appropriate literature. The project extends well beyond simply applying models to complex datasets, and thoroughly investigates a diverse range of situations, parametrizations, and sampling methods to give a very rich understanding of performance.	All key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and investigates a diverse range of situations, parametrizations, and sampling methods to give a rich understanding of performance.	Most key decisions are justified with appropriate literature. The project extends beyond simply applying models to complex datasets, and makes a good attempt to investigate a range of situations, parametrisations, and sampling methods to give a better understanding of performance.	Key decisions are justified with appropriate literature, but more depth is needed. The project extends beyond simply applying models to datasets, and seeks with some success to investigate a range of situations, parametrisations, and sampling methods to give a better understanding of performance.	Some key decisions are justified with appropriate literature, but more depth is needed. The project doesn't (or may only arbitrarily) extend beyond simply applying models to datasets; more depth of differentiated evaluation is necessary to provide a better understanding of performance.	Key decisions are not justified or substantiated with appropriate literature. The project may also lack depth or complexity in several key aspects.
Conclusion and Future Work. (10%)	Insightful conclusions, which appreciate key limitations and implications of the project. Key implications of the project are anchored with relevant literature. Well- conceived and thought out future work is discussed.	Insightful conclusions, which appreciate limitations and implications of the project. Implications of the project are anchored with relevant literature. Well-conceived and thought out future work is discussed.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Appropriate future work is discussed and presented.	Implications and limitations well understood. Discussion also correctly highlights key takeaways. Future work lacks depth and creativity, but is appropriate.	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.
	80–100	*70–79*	*60–69*	*50–59*	*40–49*	*<40*

THE FINAL MARK MUST BE 40% OR ABOVE TO ACHIEVE A PASS

CRITERIA	HIGH H1	H1	H2.1	H2.2	Pass	Fail
Quality and Presentation. (10%)	Exceptionally well written, and presented, with no mistakes in formatting or referencing. A very well-conceived video demonstrating all key functionality and the execution of key methodological aspects. The results of selected methods are illustrated with an excellent commentary that demonstrates an advanced grasp of their limitations, implications and efficacy.	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. A well-conceived video demonstrating all key functionality and the execution of key methodological aspects. The results of selected methods are illustrated with commentary that demonstrates an advanced grasp of their limitations, implications and efficacy.	Main document has a few language or style errors. Figures are well presented. IEEE template and length limit are adhered to. References are complete, and correctly used. A well-conceived video demonstrating essential functionality and the execution of key methodological aspects. The results of selected methods are illustrated with commentary that accurately discusses their implications or limitations.	Main document is readable with some language or style errors. Some figures are mostly well presented. IEEE template is largely adhered to. References are mostly complete and correctly used. A well-conceived video demonstrating essential functionality and the execution of key methodological aspects. The results of selected methods are illustrated with commentary that discusses their implications or limitations to an acceptable degree.	Main document is readable with some language 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. A demonstration video is provided that shows a functioning methodology. However, the video is poorly conceived or lacks some depth. The execution of some methods is included. Some results are shown, with little to no insightful discussion on their implications or limitations.	Littered with typos, or poor use of English. IEEE template may have been broken. Figures may be hard to read. References (if any) are probably incomplete. A demonstration video may be provided, but is poorly conceived or does not clearly illustrate key aspects of the project. Meaningful results might not be illustrated or appropriately discussed.
	80–100	*70–79*	*60–69*	*50–59*	*40–49*	*<40*

THE FINAL MARK MUST BE 40% OR ABOVE TO ACHIEVE A PASS

References

[1] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0 step-by-step data mining guide," The CRISP-DM Consortium, Tech. Rep., 2000. [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 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*, vol. 34, no. 10, pp. 895–912, 2018.