MSc Applied Data Science / Applied Data Science Degree Apprenticeship
MSc Innovative Computing / Applied Computing

APPLIED TECHNIQUES OF DATA MINING AND MACHINE LEARNING COURSEWORK

A Mini Data Mining Group Project

This coursework is the only piece for the module. It takes the form of a data mining project.

The project aims to provide an opportunity for you to gain some experience in data mining practice, but more importantly through this experience to gain knowledge and better understanding of various techniques, algorithms and solutions towards data mining and machine learning. The coursework is meant to meet the intended learning outcomes of 2, 3, 4, 5 and 6 (please refer to the module specification document in the main folder for this module on the Teams).

The project consists of three stages: project preparation, project execution and finally presentation of the project deliverables. It is very much individual contribution-based. It can be done either *individually* or as a team of *two* or *three*. Ideally, a team works better than individuals in order to create an opportunity for critical evaluation of each other's work and debating the best way forward. The scope of the project should reflect the amount of work expected from the number of people involved in the team. *Normally, each person is expected to conduct a piece of data mining from the beginning to the end using at least two data mining and machine learning methods*. Within the same application domain and problem context, each member of a team is set to either address one specific business objective or utilizing completely different approach in addressing the same objective.

Warning: please carefully control the scope of the project; you do have a proper opportunity to do a real-life individual project with a much larger and more realistic data set later in your programme. Remember that the module is worth 15 units of credit. The total number of study hours (including attending lessons) will be about 150 hours. The total number of hours on the project is about 45 - 50.

The data set, the software tools, libraries, languages and platforms for the project are entirely of your own choice. For ease of communication, members of the same team should try using the same tools and environment. For those who are comfortable with the programming, please consider using Python/scikit-learn on Jupyter Notebook. If you want to use an interactive tool, please consider Weka, RapidMiner or others.

1. Project Preparation

The first two main tasks are: (a) to form a team and (b) to select a suitable data set. Since you need time to form your team, you should perform both tasks simultaneously. You need to make up your mind almost immediately on whether you would do project alone or with someone. Please do not wait. Talk to your classmates now.

It is perfectly understandable that you may want to do the project with real-life data from your own organizations or contacts, but this is **NOT** recommended because using real-life data requires a proper project proposal and the ethical approval, which will take time. Since we only have about 8 weeks (part-time) to finish the project, it is safer to use a public domain data set of a controlled size. Please see the Reference section for some suggested data sources [1] [2], but please ensure that you read and comply with the terms and conditions for using the data.

There is NO *perfect* data set for the project. Since we are undertaking a "mini" project, please be realistic not too overambitious. A data set with thousands of rows (not hundreds of thousands) and tens of columns would be suitable. Of course, this only serves as a rough guide. If you have more people in your team, the data set can be large in terms of rows and columns.

Please complete the two tasks within *one week after* the coursework is given out. You need to inform me your team (with a nice team name?) and the data set to use.

You should then start preparing for the project by undertaking the following activities:

a) To study the project business background and understand the application domain where the data set comes from.

b) To familiarise with the process of data mining as well as the CRISP-DM methodology. We will cover the methodologies in our lecture, but you need reading into the topic.

c) To investigate the relevant literature and gain understanding on what data mining and machine learning have been used in the application domain, and what work has been done in this field. You should also identify one specific case relevant to the domain. This investigation will lead to a brief but comprehensive literature review in your project report.

d) To draw a project plan that makes sense over the working weeks within the completion date.

While you are performing the tasks above, please use your skills learnt from other modules such as the Research Methods or similar skills you already possess. Try your best to take notes in preparation for writing Background and Literature Review sections of your project report (see later). Forming a good team and working out an effective way of communication are very important too.

Advice: Since all stages of the data mining process are quite closely related, conducting a piece of data mining for a specific objective may involve repeated data preparation and pre-processing, mining and evaluation. Therefore, it is NOT recommended that the work is divided among team members according to the stages of the data mining life cycle because other members of the team have to wait while one member is working on the current stage of the work. So, each member should be running a small piece of work from the start to the end. This means that one member of a team achieves one objective and all members together achieve the overall project aim. Therefore, a good divide-and-conquer strategy is needed.

Be very much aware of the iterative nature of data mining. You may find certain patterns not good enough and decide to re-process the input data before a new round of mining begins again. Although a methodology can be followed, certain degrees of *trial and error* can never be avoided. Therefore, please allocate sufficient amount of time for the project. It cannot be done just within a few days. Therefore, you must have a good project plan.

2. Project Execution

In general, the CRISP-DM methodology should be applied and followed through the main phases of the lifecycle except the final *deployment* stage once the preparation is completed. Most generic tasks (not all of them are applicable) within each phase should be mapped to specific tasks meaningful to you own project.

You should consider an overall aim for your project upon which all data mining and data science activities are related to this aim.

Within the life cycle of the project, you may have regular discussions between the team members and make sensible decisions. Although you are working on your share of the work, exchanging "good practice" is encouraged. Constructive criticism is always beneficial to each other.

3. Project Deliverables

The deliverables of the project are in the form of a project report. Generally, you are expected to follow the CRISP-DM methodology and present your work according to the main phases of a data mining project lifecycle [3]. The report therefore should cover the following items:

a) Business Understanding. This part consists of three essential elements: (a) a background study into the problem domain, project context and the purpose of the project work; (b) the project overall aim and business objectives as well as mapping from business objectives to potential data mining tasks; and (c) a comprehensive and critical literature review that consists of a broader general review of data mining used for solving the same or similar problems followed by a more detailed study of a piece of existing data mining work relevant to the aim of this project.

b) Data exploration and understanding. This part of the report should describe the data set at hand in general, sum up any data characteristics identified through exploration of the data set, plus a critical evaluation of the data quality.

c) Data Preparation and Pre-processing. This part of the report should present a set of pre-processing operations conducted in order to prepare the data for the modelling stage. The work conducted at this stage should be backed up by valid reasons. You can include multiple versions of pre-processing to reflect the repeated nature of the project.

d) Data Modelling and model evaluation. This part of the report documents the details of the mining tasks conducted and their outcomes. The evaluations of any discovered patterns or models are also presented here. Some interpretation and further analysis of the outcomes are also described here. Any further post-processing tasks should also be documented and supported. Just a piece of advice here: you might conclude that the discovery operations so far are still insufficient for finding the final and useful truth. In other words, you may have a feeling of quite "open ended" and you could continue and go on further. This is quite common and natural feeling to have for data science projects. So if you have already done a substantial amount with sufficient depth, you should stop here, and wrap up any possible ideas for further trials as "Future Work" in the Summary part next.

e) Project Evaluation and Summary. This is the final part of the report. It summarizes the work of the whole project and highlight any major findings through the project. You can recommend either the need for another round of data mining to be done or a set of possible deployment actions. Any parts of the project that could be expanded into can be presented as future work. You should also highlight potential impacts of the project results towards the business and/or society.

4. Project Assessment

The project is valued not really about what have been found from the data, but more importantly about how the project has been conducted, and what you learn from the project experience. This is the key difference between this coursework module project and your final individual project in data science. A specifically designed rubric will be used to determine the mark grades. It will be published in due time.

The outline of the mark breakdown is as follows:

(a)	Business Background Understanding:	20
(b)	Data Understanding and exploration:	15
(c)	Data Preparation and Pre-processing:	25
(d)	Modelling and model evaluation:	25
(e)	Project Evaluation and Summary:	15

For data mining projects, it is not quite easy to specify a word count or a maximum page number for the report. At the same time, we do need to constrain the size of the report to reflect the scope of the work. Therefore, the whole report should not go significantly beyond 4,000 words. Some selected charts, tables and figures should appear in the main body of the report. Any useful supplementary materials should be included in the report appendices.

This project counts for 60% of the total mark for the module. The submission date will be Monday Week 9.

Hongbo Du
Module Leader

References

[1] UCI Machine Learning Data Repository, https://archive.ics.uci.edu/ml/index.php, accessed on 07/03/2022

[2] Kaggle Open Data Sets and Machine Learning, https://www.kaggle.com/datasets, accessed on 07/03/2022

[3] CRISP-DM Methodology, https://www.datascience-pm.com/crisp-dm-2/, accessed on 1/4/2021

[4] Hongbo Du, "Data Mining Project: A Critical Element in Teaching, Learning and Assessment of a Data Mining Module", BNCOD2011: Advances in Databases, Lecture Notes in Computer Science (LNCS) Volume 7051, Springer, (downloadable from: https://www.researchgate.net/publication/220862936_Data_Mining_Project_A_Critical_Element_in_Teaching_Learning_and_Assessment_of_a_Data_Mining_Module)