# Assessment 2: Mining a dataset

**Value: 35%**

**Due dates:**

Demo: scheduled labs on the week of December 8th

Report: Friday December 12th, 6pm

**Objective**: To mine a dataset using the CRISP-DM methodology.

**To do:** Each student has been allocated a dataset on moodle. Mine the dataset following the CRISP-DM methodology. Write a report on the work done during each phase of CRISP-DM. A report template is available on moodle, and further details on each phase is given below:

Business Understanding:

Give an introduction to the dataset, and state the Business Objective and Data Mining Objective(s) for the mining project.

Data Understanding:

Describe the attributes in the dataset in terms of quality, information content, and usefulness for the mining objective using EDA techniques.

Data Preparation:

Decide on at least three data preparation techniques to use. Take a look at the workflow on Moodle for ideas.

Discuss why your chosen techniques are appropriate/required for this data set and mining objective.  Document the improvements, if any, on the accuracy of the mining result as a result of each step in the data preparation phase. For example, if choosing sampling, discuss how you arrived at the optimal sample size, and discuss the accuracy levels achieved for different sample sizes justifying your final sample size.

Data Mining:

Use at least two mining algorithms on the dataset.

For each algorithm: Justify why this is a good algorithm to use on the dataset, and find the optimal parameter values for your mining algorithms.

*Note: Because of the iterative nature of data mining, the data preparation and data mining sections can be combined in your report if you find that more convenient.*

Evaluation:

Discuss the overall accuracy of your final model, and explain, in non-technical terms, what you have learnt from the dataset.

**What to submit:**

1. All students must do a demo of their work in RapidMiner (or R) during the scheduled labs on Dec 8th and 12th. The objective of this demo is to verify the work submitted is your own. Failure to complete the demo will result in 0% for this CA.
2. Upload your report to Moodle/TurnItIn as detailed above.

**Project Plan:**

| **Task** | **Completed by** |
|---|---|
| Business Understanding and Data Understanding – 1 week | To be completed in labs during Nov 10th – 14th |
| Data Preparation – 2 weeks | To be completed in labs by Nov |

| | |
|---|---|
| | 24th – 28th |
| Data Mining & Evaluation – 1 week | To be completed in labs by Dec 1st to 5th |
| Final demo of work done & upload report to moodle | Demo assessment work during scheduled labs, Dec 8th and 12th. Upload your final report to moodle by Dec 12th, 6pm. |

**Marking scheme:**

Business and Data Understanding: 10%

How well do you understand the mining objective, and how well do you understand the quality and information content of the dataset?

Data preparation: 15%

Were your chosen techniques appropriate for the dataset? Did you apply them properly and understand the impact of each technique on the overall accuracy of the mining algorithm?
Is there evidence of experimentation (e.g. tried something, evaluated it but it didn't seem to make much difference, had another idea to try . . . . .).
This phase is about experimenting with different ideas that make sense for the dataset you are working with.

Data mining: 15%

Were your chosen mining algorithms appropriate for the data-mining task? Did you understand how to use the algorithm and interpret the results? Did you understand the parameter settings for the algorithm?

Evaluation: 5%

Were you able to explain, in non technical terms, what you learnt about the dataset, what could be predicted from it, and what, if any, were the limitations / inaccuracies.

Overall quality of the report: 5%

Was the report well written and well laid out

**Including any plagiarised work in a submission will result in 0% for the submission. Repeated instances of plagiarism will result in 0% for the module.**