**Data Analytics Project: Mining a dataset**

**Value: 30% of the overall grade (the score is first out of 100% and it will then be weighted by 0.3)**

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

**TO DO:**

1. Analyse a dataset using the CRISP-DM methodology and then write **an interim report** (on the first 2 phases: Business Understanding and Data Understanding; not graded, but essential for peer review) and **a final report** on the work done during each phase of CRISP-DM (90% of the overall project).
    - You will be allocated a dataset to mine (see allocation on Moodle).
    - A report template is available on Moodle.
2. Complete a **peer review** (10% of the overall project).
    - You will be allocated a project to review (see allocation on Moodle).
    - An evaluation form is available on Moodle.

**FURTHER DETAILS ON EACH PHASE AND MARKING SCHEME ()**:

## 1. Report (90%)

Business Understanding 6%:

- Give a brief background to the domain and dataset.
- State the Business Objective(s).
- State the Data Mining Objective(s) for the mining project.

*Question to ask yourself in this section: How well do you understand the task objectives? Can you translate the business objectives into suitable data mining objectives?*

Data Understanding 20%:

- Describe the dataset in terms of dimensions, and what columns are used as regular attributes and what column is used as the label column.
- Discuss interesting/important attributes in the dataset in terms of quality, information content, and usefulness for the mining objective using EDA techniques.
    - Derive and interpret summary statistics, and use at least 3 types of appropriate charts in your report. It should be clear from your report that you can interpret summary statistics and charts and understand their relevance to preparing a dataset for the mining algorithms.
- Identify any quality issues and discuss potential problems that they may cause.

*Questions to ask yourself in this section: How well do you understand the task objective, and how well do you understand the quality and information content of the dataset? Did you use a variety of statistics and graphs and interpret them correctly?*

Baseline and Test Designs 10%:

- Decide on one classification algorithm that will subsequently be used to record the performance after each technique applied throughout the entire project; you must document this choice and explain why.
  - it should be based on the dataset properties (e.g., is it suitable for the size of your dataset, does it produce an easy to interpret model, etc.).
- Decide on the test design and appropriate performance metrics and justify why you chose them:
  - E.g., use cross-validation/split validation and use metrics such as accuracy, precision, and recall/f-measure.
- Generate the baseline model and document and discuss its performance.

*Questions to ask yourself in this section: Did you choose an appropriate baseline algorithm, appropriate metrics, and appropriate test design?*

Data Preparation 25%:

- Decide on at least 3 suitable data preparation techniques to use so that it addresses quality issues, as well as helping the mining algorithms identify patterns in the dataset. Justify the choices made, i.e., discuss why your chosen techniques are appropriate/required for this data set and mining objective. It should be clear from your report that you understand each technique, and why it is used.
  - For example, if choosing sampling, discuss how you arrived at the optimal sample size, and discuss the performance levels achieved for different sample sizes justifying your final sample size.
- Document and discuss the performance after each preparation technique is applied, i.e., you must re-apply the chosen classification algorithm and get the performance to understand whether it improves, disimproves, or stays the same, and based on this, you should decide whether to include it in the final pipeline/model. It should be clear from your report that you understand the potential impact of each technique on the classification model performance.

*Questions to ask yourself in this section: Were your chosen techniques appropriate for the dataset and informed by the data understanding phase? Did you apply them properly and understand the impact of each technique on the overall performance 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 techniques that are relevant for the dataset you are working with.*

Hyperparameter tuning 15%:

- Each classification algorithm has a number of parameters that you can adjust:
  a) experiment with different parameter settings applicable to your chosen classification algorithm and discuss what they do.
  b) document the performance for various ranges and combination of parameter setting tried and comment on how they affect performance, and whether they lead to overfitting or underfitting.

*Did you understand algorithm parameters, and experiment with different parameter values? Did you understand how to interpret the results?*

Evaluation 10%:

- Discuss the overall performance of your final best model:
  - What preparation techniques worked best?
  - What were the optimal parameters values?
- Explain, in non-technical terms, what you have learnt from the dataset. For example, what attributes are predictive and what attributes are not, did this concur with your initial understanding of the attributes? Did you find any surprising facts?
- Also discuss any limitations of the dataset that may have affected the model performance and how would you potentially address them.

*Questions to ask yourself in this section: Were you able to explain, in non-technical terms, what you learnt about the dataset, what could be predicted from it, what variables were predictive and why, and what, if any, were the limitations / inaccuracies of the mining process, etc.?*

Report Layout/Write-up and References 4%:

- You are expected to use a coherent/ logical flow and good English.
- Any websites, books, articles, lecture notes, colleagues/friends you have used to complete the project must be referenced.

Appendix:

- Copy all your code in this section – this is also a must.
- Note that this section will be exempt from plagiarism check.

2. **Peer Review 10%** (see template provided on Moodle).
   - Review and provide feedback on the business understanding phase of the allocated project.
   - Review and provide feedback on the data understanding phase of the allocated project.

**WHAT TO SUBMIT:**

1. Upload your **interim report** to Moodle using the designated facility by 3rd November (11.55 pm is the deadline).
2. Upload your **peer review evaluation form** to Moodle using the designated facility by 10th November (11.55 pm is the deadline).
3. Upload your **report** to Moodle using the designated facility by 8th December (11.55 pm is the deadline).

**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.**

**All instances of plagiarism will be recorded with the Course Coordinator.**

**MARKING RUBRIC FOR REPORT**:

| Phase/Stage | Weak (F - D) (lack of understanding / evidence of understanding is not clear in any of the areas assessed). | Average (C – B) (evidence of some understanding in most areas assessed / good understanding is some of the areas assessed) | Strong (B+) (evidence of good understanding in most/all areas assessed) | Excellent (A) (clear evidence of excellent understanding in most/all areas assessed) |
|---|---|---|---|---|
| Business Understanding | Not included/ irrelevant background and no separation between business and mining objectives. | Some background provided; both business and mining objectives, but the mapping between them is unclear. | Sufficient and relevant background. Clear business objectives well mapped to mining objectives, but mining objectives lack technical detail. | Sufficient and relevant background. Clear business objectives well mapped to mining objectives, and the mining objectives technically detailed. |
| Data Understanding | Relevant summary statistics and charts are not included / included, but not interpreted / wrongly interpreted / not relevant to the mining objectives. | Summary statistics and charts are included, but only some are interpreted and/or only some are relevant to the mining objectives. | Relevant summary statistics and charts to the mining objectives are included, and interpreted correctly; however, some explanations lack detail. | Relevant summary statistics and charts to the mining objectives are included and interpreted correctly and in sufficient detail. |
| Baseline & Test Design | Not included / included, but no justification and explanation of algorithm, test design and metrics / justifications&explanations are not appropriate; performance not included / included but not discussed. | Included, but not all are justified and explained / some justifications&explanations are not appropriate; performance included but not discussed. | Included, and all are justified and explained appropriately, however, with limited detail; performance included and discussed. | Included, and all are justified and explained appropriately, in sufficient detail; performance included and discussed. |
| Data Preparation | Not included / included, but no justification and explanation of preparation techniques / justifications&explanations | Included, but not all are justified and explained / some justifications&explanations are not appropriate; performance included but not | Included, and all are justified and explained appropriately, however, with limited detail; performance included and | Included, and all are justified and explained appropriately, in sufficient detail; performance included and discussed. |

| | | | | |
|---|---|---|---|---|
| | are not appropriate; performance not included / included but not discussed. | discussed. | discussed. | |
| Hyperparameter Tuning | Not included / included, but no justification and explanation of hyperparameters tried / justifications&explanations are not appropriate; performance not included / included but not discussed. | Included, but not all are justified and explained / some justifications&explanations are not appropriate; performance included but not discussed. | Included, and all are justified and explained appropriately, however, with limited detail; performance included and discussed. | Included, and all are justified and explained appropriately, in sufficient detail; performance included and discussed. |
| Evaluation | Not included / wrong conclusions. | Brief overview of results with limited discussions. | Good overview of results but discussions lack detail in one or two areas. | Good overview of results; adequately detailed discussions. |
| Layout/Write-up | Messy layout/flow with poor quality English. | Good layout/flow, but with poor quality English / messy layout/flow with good quality English. | Good layout, mostly coherently structured and good quality English. | Coherently structured report with excellent quality English. |