**MIS772 – Predictive Analytics**

T2 2024

Assignment 1 – Individual

Student name: Type your name here

Student number: Type your student number here

# **Executive summary (1 page)**

Use this section to include an executive summary written for a senior business manager/executive or similar non-technical reader. In this section, use professional business language and not computer / mathematical / statistical / data science /conversational language/terminology. Address the following:

**Executive problem statement:**

* Describe the problem in clear terms from the business perspective and not a technical perspective.
* State the high-level objective(s) and not the methods of their achieving.
* Ensure that the problem you describe can be solved using the provided data.

**Executive solution statement:**

* Answer each of question(s) posed to you by the business.
* Clearly state the recommendations/decisions to be made/actions to be taken from your analysis and modelling.
* Do not include any charts or tables in this section. However, include cross-references to specific analyses and results in the subsequent sections of your report, e.g. “… (see Figure 1)” or “As shown in Table 4…”.
* All aspects of the solution and recommendations must follow directly from your analysis/modelling. Do not include any speculative/unfounded statements/unsubstantiated claims.
* State the anticipated benefit(s) of the solution for the company.

**Rubric reference:** Executive summary of problem statement and solution

**All comments, such as this, which are not part of your submission should be deleted to save space.**

# **Data exploration, pattern discovery, and preparation (2 pages)**

Write this section and all subsequent sections of your report for an informed technical/business analyst in the organisation. In these sections you can use computer / mathematical / statistical / data science language/terminology as appropriate.

Use this section to include all your work on exploring the data set and preparing it for further processing in RapidMiner. Also, use this section to include all your work on finding meaningful patterns in the data set as relevant to the case study using RapidMiner. This section may include:

* your selection of relevant data attributes as predictors
* your selection of an attribute as label
* your ways of dealing with missing values
* transformations you are doing on the original data (can be any textual modifications on the data, any numeric normalizations, or any type conversions such as nominal to numeric or similar)
* your data distribution analysis (e.g., using histograms or scatterplots)
* your correlations analyses using correlation matrices

Make sure your visualizations are accompanied by relevant discussions of the insights the analyses and visualizations will/should lead to. BRING IN SCREENSHOTS OF YOUR RAPIDMINER PROCESSES AND EXPLAIN THE MAJOR FUNCTIONALITIES…

* If you include any charts or tables you must describe them (e.g. by using arrows / boxes).
* Make sure that any included chart is readable (so do not shrink it into microscopic size).
* If you scale the included screen shots keep their proportions (do not distort images).
* Most importantly describe what those data features mean and how important they are, and why
* Avoid indiscriminate “dumping” of tables, charts or code into this section – all content must have a purpose relating to the problem statement.
* All included charts, tables or RM processes (or their parts) have to be described or used in the discussion.
* Make sure that all charts, tables and important results are labelled for cross-referencing, e.g. “Figure 1 - Histogram of Overall Rating”

**Rubric reference:** Explore, discover, and prepare data for predictive analysis

**All comments, such as this, which are not part of your submission should be deleted to save space.**

# **Predictive modelling (2 pages)**

Use this section to show and discuss the details of RapidMiner process(es) to develop predictive models to address the problem and specific questions.

This may include:

* k-NN classification(s), Decision Trees (or forests), or other classifiers covered in this unit.
* your justification of why you chose to that specific modelling

BRING IN SELECTED SCREENSHOTS OF YOUR RAPIDMINER PROCESSES AND EXPLAIN…

* Describe processes, operators and their parameters (annotate with text and arrows).
* Justify the values of the model parameters
* If your process is very large, consider splitting it into sub-processes or separate processes, or include only the most important parts.
* Utilise cross-referencing with tables or charts from other sections if needed.
* Avoid indiscriminate “dumping” of RM processes/models into this section – all content must have a clear purpose.
* Do not include definition of terms or a “textbook” description of operations – this is already known!

**Rubric reference:** Analyse data with predictive models

**All comments, such as this, which are not part of your submission should be deleted to save space.**

# **Model evaluation and improvement (2 pages)**

Use this section to report and explain the performance of developed classification models in RapidMiner. Also, report any steps you have taken to improve the performance of your model(s).

* your evaluation procedure (e.g., hold-out or resampling) for any of the analysis cases you have included in the previous section and results (e.g. using accuracy and kappa)
* your comparative analysis on the evaluations of different predictive models you have created (e.g., decision trees as well as k‑NN)
* any performance improvements you may have achieved by trying different evaluation procedures (e.g., hold-out vs. resampling)
* any performance improvements you may have achieved using ensemble techniques (e.g., voting, stacking)
* You need to describe and explain your model evaluation results.
* If you have few results to report, include here screenshots of your results, e.g. confusion table or ROC charts.
* If you have many results to report, include here a table of all results.
* Explain the impact of the obtained results on the future use of the model to support decision making.
* Avoid indiscriminate “dumping” of performance results – all content must have a clear purpose.

**Rubric reference:** Evaluate and improve analytic solutions

**All comments, such as this, which are not part of your submission should be deleted to save space.**

**Any materials, analysis or reports that do not fit into 7 (seven pages in total) will not be assessed or marked.**