You are required to access a large data set and apply the CRISP-DM methodology to meaningfully clean, transform, analyse and evaluate it. As part of this process, you are required to subsequently apply one or more machine learning technique(s) of your choice to perform classification, association, numerical prediction and/or clustering tasks (or combinations thereof).

You will present the outcome of the above tasks in the form of a technical report containing the five sections listed in Table 1.

Section	Weighting	Recommended Pages per Section
1. Introduction and Business Context	0.1	1
2. Data Selection and Pre-Processing	0.3	3
3. Machine Learning Method(s) and their Implementation	0.3	3
4. Evaluation of Results	0.2	2
5. Discussion	0.1	1

As shown in Table 1, a page limit of 10 pages is recommended. The report, in total, however, must not exceed 13 pages (excluding title page, contents page, references, bibliography and appendices) with a minimum font size of 10 pitch. A penalty of a single grade will be incurred if you exceed the 13-page limit. Further information (supporting experimental results) can be added as appendices.

You are free to select the style of the report (i.e., section headings and format, etc.) although it must obviously address the content listed in Table 1.

You are expected to submit the following electronic files to the designated repository by the submission deadline:

- **Training, validation and test sets (before and after pre-processing). Note that if cross-validation is used, only the training and test sets are required;**

The remainder of this section provides you with detailed requirements for each area of content.

PROJECT TITLE

Table of Contents

1. Introduction

You should also look at how to use automatic referencing and citing (Harvard referencing style). You may want to create 'sections' (e.g. Insert ->Break->Section break /continuous or next page) so that you can apply different formatting to different sections of the document. You should make use of sub-section headings (via relevant Styles) where appropriate Section 1 includes:

▪ A brief narrative as to how your organisation or company currently performs data analytics and how the CRISP-DM methodology may help your organisation to better meet its strategic priorities with respect to data analytics and business intelligence.

▪ A brief overview of the data analytic task you are going to perform.

▪ A clear justification as to why the task you are attempting is of value to your business or more broadly, industry, government, university research and/or the community. You should support your justification with references to the appropriate industry or academic literature.

▪ State the insight you intend to gain

2. Data Selection and Pre-processing

● Select a data set consisting of at least 2,000 observations/records and preferably above 10,000. You are strongly encouraged to identify an anonymised data set, the strategic objectives of the business. However, if this is not possible then you are advised to select a data set from one of the following sources:

https://archive.ics.uci.edu/ml/datasets.html

https://www.kaggle.com/datasets

http://www.cs.waikato.ac.nz/ml/weka/datasets.html

http://yann.lecun.com/exdb/mnist/

https://www.springboard.com/blog/free-public-data-sets-data-science-project/

● Briefly describe your data set and reference its origin.

● If you have 15 or less attributes, table your attributes with attribute name, description and data type and then show the minimum/average/maximum and stdev values for the training set and test set. For nominal variables, then show the most and least frequently occurring nominal value(s). If you have more than 15 attributes, then group attributes into themes (e.g. customer, orders, employees) and describe the type of information and data types in each theme including number of each variable type (e.g. nominal, interval, ratio etc). You may want to highlight significant variables identified by some attribute selection algorithm.

● Briefly table the following characteristics of the entire data set: number of instances, patterns per target class (if classification), limitations such as possible conflicting patterns, missing values, outliers/erroneous values.

● Explain how you have sampled your data to create the 'in sample' and 'out of sample' data sets. If you have used instance weightings to balance your data set(s) then explain how the weightings were determined.

● Provide a statistical summary in tabular form for the resulting 'in sample' (training/validation set) and 'out of sample' (test set). Also, state whether or not there was any overlap in training and test set instances and if so, justify why your test set is not compromised.

● What pre-processing and transformation was performed on the variables and why? (e.g. standardising numerical variables and/or using scaling, taking logs to reduce skewness, or log differences to reduce non-stationarity; converting numerical variables to discrete ones; converting numerical or symbolic patterns into bit patterns; removing patterns with missing or outlier values; adding noise or jitter to patterns to expand the data set; adding instance weightings or replicating certain pattern classes to improve class distributions; transforming time-series data into static training/test patterns)

● How did you ensure that your pre-processing did not compromise your test set (e.g. use of standardisation)

● Consideration will also be given to the 'curse of dimensionality', its issues and how its impact can be reduced. o If you reduced the number of dimensions (e.g. from 30 attributes to 10 attributes), how did you do this? Autoencoder? PCA? Filter using InfoGain measurement? A clusterer? How do these methods work and what are their advantages/disadvantages?

● If you increased the number of training instances, how did you do this?

3. Machine Learning Method(s) and their Implementation

▪ Clearly state the machine learning methods you will be using the function(s) you will be expecting them to perform (e.g. classification, association, regression, clustering or combinations thereof for self-supervised learning). You must describe the expected 'input to' and 'output from' each model.

▪ Explain and justify the machine learning method(s) chosen for the task. You must also use a simple benchmark model with which to compare your chosen machine learning model(s) (e.g.

benchmark a neural network trained with back-propagation against a simple OneR or Naive Bayes approach).

▪ Briefly highlight the strengths and weaknesses of the chosen learning method(s).

▪ Describe your 'model fitting' and 'model selection' process (e.g. leave-one-out validation, cross-validation, bagging and boosting etc). You must state and justify the hyper-parameters used for model fitting and how 'over-training' will be minimised.

▪ Describe what tool will be used to implement the machine learning method(s) (e.g. Weka/Java).

▪ You must either: a) use advanced features of the chosen analytics tool including (though not limited to) clear evidence of meaningful programming/scripting activity to use machine learning and/or pre-processing tools in a bespoke way (e.g. install and use advanced Weka packages via Package Manager – examples might be: simple recurrent network, convolutional neural network, Self-organizing maps, Time Series processing with ARMA models).

OR ▪ provide an in-depth mathematical treatment of the chosen machine learning method(s) with clear explanation as to how you will optimise them using the built-in features of the data analytics tool.

4. Evaluation of Results

▪ Table the resulting 'in sample' (training) and 'out of sample' (test) performance of your model for the different model configurations and trial runs (e.g. a neural net with different number of hidden nodes, different random starting weights and or different learning rates). You should at least use one or more of the performance metrics (as appropriate):

o Percent correct/incorrect

o Confusion matrix

o Recall and precision

o Evaluating numeric prediction (e.g. mean squared error (MSE), root mean squared error (RMSE), correlation coefficients)

o ROC curve

▪ Critically review the performance of the different models. Which type of pre-processing appeared to be most advantageous and why? For each model, which hyper-parameter settings (e.g. learning rate, prune Tree, momentum term) were most effective?

- ▪ Critically compare models – was there a model or model class whose performance on the test set was statistically significantly better than the other models/model classes (with a p-value < 0.05) (may be using Experimenter in Weka)?

5. Discussion

▪ Briefly summarise your task and your findings (i.e. whether the model learnt the problem).

▪ How do your findings relate to similar tasks found in the relevant industry or academic literature?

▪ Did you gain the insight you intended to? If not, what else could you do to enhance the usefulness of your analytics?

▪ How did you decide on the most appropriate machine learning method and what do you understand about appropriateness?

▪ Finally, briefly state how you are going to use the knowledge and skills you have developed in the module to further your professional ambitions and/or the strategic objectives of your organisation?

References

Harvard referencing style. Use the MS Word citing and referencing feature

A. Appendix

This is an example of an appendix containing additional information you might want to show:

e.g. initial experimental results before identifying optimum parameter

Assessment Criteria

Section 1 Introduction: A cogent and engaging account is given of the data analytic task and its broader relationship/impact on business, industry and society. A concise critical account of the organisation's approach to data analytics given with respect to all six phases of CRISP-DM (with additional material in the appendices). Justifications provided are supported by the literature. A deep understanding of the problem area and industry affected is clear from the discussion provided and potential value/insight expected.

Section 2 Data Selection and Pre-processing: A substantial and meaningful dataset has been selected. There may be evidence of bespoke modifications. A critical explanation and statistical summary are provided with references to industry/published work which has processed similar data sets. All of the relevant preprocessing requirements of the spec have been addressed with critical consideration and justification of the methods used (e.g. removal of outliers or seasonality for regression analysis, standardised numerical variables, and simplified task by converting continuous data to discrete or symbolic form). For higher grades, some dimensionality reduction was used (e.g. WEKA's Gain Attribute Evaluation filter or similar). The organisation of the data into training, validation and test samples is well articulated and justified by the literature.

Section 3. The Machine Learning Method and its Implementation: An appropriate benchmark model has been identified along with a number of more powerful machine learning models (some not taught such as Random Forests). A critical discussion is given as to how it will be applied with

references made to relevant research papers. A concise and erudite discussion is given as to the model fitting and selection methods that will be used. Bagging and boosting have typically been considered. A deep understanding is evidenced also through reference to the 'curse of dimensionality' and 'over-fitting' and how to minimise issues. For higher grade, a strong attempt has been made to use advanced features of tool or to describe models mathematically with accuracy. If WEKA used then additional packages were installed and successfully used for Kohonen clustering, simple recurrent networks, or timeseries processing (ARIMA models), etc.

Section 4. Evaluation of Results: It is evident that a clear & wellstructured set of experiments was performed to identify optimal preprocessing methods and training parameters. All key training and test results are clearly and professionally tabled using one or more appropriate measure (RMSE, MSE, Precision /Recall, F-score, % correct, ROC) including those not taught (e.g. Kappa statistic, DickeyFuller test). Also, the trained models have been selected based on some holdout validation set or cross-validation. The results are critically discussed with respect to a benchmark and clear references to the literature. Higher grades: Ensemble methods, e.g., bagging and boosting MUST have been applied to produce classification/predictions from an array of models. Statistical hypothesis testing MUST have been applied to justify significance of results (i.e. t-tests, p-values) and whether one model is statistically significantly better than another.

Section 5. Discussion: A clear and concise summary of the work and findings is given. A critical and introspective account is given and indicates the success of the data analytics project with some meaningful reference to published literature (and organisations goals). An erudite account of the insight gained is provided together with clear indications as to how to advance the work further.