

Group Assignment:

This assignment is a group assignment consisting of 2 people. Students are responsible for the formation of groups.

Introduction & Overview:

The purpose of the assignment is to use RapidMiner to apply the concepts you have learned about the Data Analytics process to a **public** dataset of your choice e.g. from the UCI repository or Kaggle. Students are gently reminded that due to GDPR regulations and subtleties concerning “informed consent”, “retention”, “security”, “transfer”, “privacy” etc. it is unfortunately not possible to use private datasets.

Required Tasks:

You are required to produce a report detailing your work investigating the data, building a Data Analytics model, analysing the results and comparing your results with published findings in the area.

The first task you should complete is a data exploration exercise, where you will document the summary characteristics of the data set and further understanding that you have gained through visualisations.

Depending on the problem domain, develop a supervised or unsupervised model of the dataset e.g. pick a candidate technique after researching standard approaches to the problem in the published literature. The modelling phase should be **iterative** i.e. adjust the settings of the model until you are satisfied that the performance is acceptable.

You will need to independently evaluate the results of the model on a separate test set. You can then compare your results with published findings and discuss the outcomes.

Deliverables:

The report should be roughly 5 pages long and not longer than 8 pages. A separate index should be included at the end of the document for any references cited and an appendix for additional relevant material generated as part of the process. The complete document should not exceed 10 pages.

The report should clearly show your work in the following areas:

- Definition of problem.
- Data Exploration and Descriptive Analytics.
- Identification of data insights from previous step.
- Details of the model building iterations conducted e.g. intermediate results, rationale for adjusting configuration settings etc..
- Details of the evaluation strategy and performance measures for your data analytics model.
- Discussion of reasons underlying how your results compare and contrast with 1.) your expectations after selecting the final model configuration settings and 2.) further results for the dataset that are available in the online literature.

It is fine to include a selection of important tables and graphs in the main report (other tables and graphs created and referenced in the report should be captioned and put in an appendix). It is **not required** to include screenshots of RapidMiner processes in the report as these are readily available in the process files. The file SampleAssignmentDataAnalytics.pdf contains an example of a report previously submitted for this assignment.

Individual Work

My experience is that it can be the case that students treat assignments as box-ticking exercises to be completed, submitted and forgotten about. Such an attitude lets slip the opportunity project work represents to develop relevant skills and experience that can be used as a platform to build on going forward e.g. the recent job descriptor, [ExampleJobDescriptor.pdf](#), included in the zip file indicates that employers are usually very interested in the attitude of the applicant and evidence of how this can be authentically demonstrated.

The last part of the assignment requires that each student individually defend their role in the realisation of the project. The format for this is 1 page of text per student, written in the style of answering the following question at a job interview “So, could you tell us a little about project work that you were involved in as part of the Masters in Data Analytics”.

Submission Details:

The assignment is due by Friday 16th of December 2022 at 23:55. You will need to submit your assignment via the link available on Moodle.

The submission format is a zip file containing:

- 1.) 1 jointly authored report in the form of a Word processed document i.e. .doc, .docx file or OpenOffice equivalent (note pdfs will not be accepted).
- 2.) A 1 page individual write-up defending their role in the realisation of the project.
- 3.) 1 RapidMiner process containing the DM pipeline and anything required to run it without error on the examiner's machine.

Marking Scheme:

The marking scheme for the assignment is as follows:

- 15% Problem Definition, Descriptive Analytics, Data Insights, etc & summary of initial findings/insights.
- 15% Details of the iterations undertaken in building the model.
- 15% Details & Discussion of the evaluation and performance measures for data model and comparison with existing work.
- 30% Complexity of the work undertaken in terms of dataset chosen, originality of approach and depth of understanding of the data mining process demonstrated.
- 25% 1 page individual write ups.

Each submission must be original work as [plagiarism](#) will result in a zero mark (0%). There will be a 10% penalty deduction applied for each day the assignment is late. There is no penalty for submitting early!