# University of Sunderland

**School of Computer Science**

**CETM72 Assignment 2 – Implementation and report**

**Data Science Principles**

**This assignment contributes 65% to your final module mark.**

The following learning outcomes will be assessed:

1. Critically select and apply key machine learning and statistical techniques for data analytics projects across the whole data science lifecycle on modern data science platforms and with data science programming languages.
2. Appropriately characterize the types of data; to perform the pre-processing, transformation, fusion, analysis of a wide range type of data; and to visualize and report the results of the analysis of various types of data.

**Important Information**

You are required to submit your work within the bounds of the University Infringement of Assessment Regulations (see Programme Guide). Plagiarism, paraphrasing and downloading large amounts of information from external sources, will not be tolerated and will be dealt with severely. Although you should make full use of any source material, which would normally be an occasional sentence and/or paragraph (referenced) followed by your own critical analysis/evaluation. You will receive no marks for work that is not your own. Your work may be subject to checks for originality which can include use of an electronic plagiarism detection service.

Where you are asked to submit an individual piece of work, the work must be entirely your own. The safety of your assessments is your responsibility. You must not permit another student access to your work.

Where referencing is required, unless otherwise stated, the Harvard referencing system must be used (see your Programme Guide).

Please ensure that you retain a duplicate of your assignment. We are required to send samples of student work to the external examiners for moderation purposes. It will also safeguard in the unlikely event of your work going astray.

|  |  |
| --- | --- |
| **Submission Date and Time** | As advised on Canvas |
| **Submission Location** | Via Canvas |

**Your Task**

**THIS ASSIGNMENT REQUIRES R CODING AND A SHORT REPORT (65% of module marks)**

Your task is to conduct data analysis on a given data set from the UCI site. To help you in this task please look over our past RStudio activities where we loaded in data, pre-processed it, trained machine learning algorithms on the data and plotted the results.

The first part of the report is simply text describing the introduction, application area and data to be used, machine learning algorithms to be used.

What I expect to see for the practical implementation part of the report are screenshots of your code in the RStudio script editor. Screenshots of key outputs and screenshots of important diagrams. Along with text to describe what I’m seeing and identify any salient points. The presentation of your practical work should be identical to the way I’ve presented the Activities in R over the last seven weeks. You need to use snipping tool in Windows or similar to grab screenshots of selected areas.

Finally, write up your work in a 1,500 word (+/- 10%) report

**Student Information**

The report should include the following headings:

**Report – (40 marks)**

Introduction (10 marks)

Application area and data (10 marks)

Machine learning algorithms (10 marks)

Conclusion, structure of report, including refs (10 marks)

**Practical Implementation - (60 marks)**

Pre-processing on real or simulated data (10 marks)

R Programming content and your function (20 marks)

Display of data/results (20 marks)

Source code listing (10 marks)

State the R packages you have used, any source code you have used from others. Also, place a full R source listing at back of report - it will not add to word count but DO NOT go over page count of 15 pages

You can refer to any of your course handouts, any other books, journals, online resources etc.

1. **Introduction**

Your introduction should include a summary of the main points that you will discuss in your report. Your report should outline the area your data is from and what you hope to achieve. Your introduction should be about 150 words in length.

**2. Data used**

The purpose of this section is to ensure you understand the types of data and the pre-processing you will use. What types of variables are present such as: integer, dates, strings, etc. Provide literature and examples associated with your data set. This section should be approximately 150 words.

1. **Machine learning methods used**

In this section you should identify the machine learning methods that you will apply to the UCI data. What criteria will be used to measure the success of the machine learning methods. This section should be approximately 150 words.

1. **Practical: Pre-processing of data**

In this section you should discuss how the data was read in, what pre-processing if any occurred and why you did it. Show me screen shots of code with your text write up. This section should be no more than 150 words in length.

1. **Practical: R Programming content**

In this section you should show me screen shots of code with your text write up. The R programming content can include building your machine learning models, testing of models, perhaps you have done a compare/contrast with several models. I would also like to see an R function written by you. The source code should be neat and tidy, use comments where necessary to explain the main actions of your code. This section should be no more than 300 words in length.

1. **Practical: Display of data/results**

This section you should use screenshots of key R output, important diagrams and anything to do with your machine learning models. Along with text descriptions of the outputs. It should be no more than 300 words in length.

1. **Source code listing**

This includes all your R code including the library commands. I expect to be able to load in the libraries you have used and copy and paste and run your analysis.

1. **Conclusions**

In this section you should summarise your experimental results and findings. This section should be approximately 150 words.

1. **References and look and feel of report**

These should be to Harvard standards (not included in work count but should be between 5-10 references). References should be valid and appropriate. The formatting of the report should be neat and tidy. Diagrams should be used with good descriptive text. Diagrams should be easy to read, and a sensible number of no more than 6-7 diagrams used. **No more than 15 pages in total for everything including source code listings, put source code listing in font size 10.**

The word counts for the sections are just advisory based on marks allocated.

**Submission Guidelines**

Your report should be spell checked and contain references. You must use the Harvard style of referencing, both for citations within the text and your reference list. It is important that you read thoroughly the information on the cover sheet regarding the university assessment regulations, including those regarding plagiarism and collusion. Assignment hand-in requirements are specified on the front cover sheet. The approximate time you should spend on this assignment is 30-50 hours. Your assignment must be handed in before the time specified. Your assessment will be assessed according to the University’s Postgraduate Generic Assessment Criteria, which are provided on the following pages.

**CETM72 Data Science Principles – Assignment 2 Marking Sheet**

**Name Student Registration Number**

**Total Mark %**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Categories** | | | | | | |
|  | **Grade** | **Relevance** | **Knowledge** | **Analysis** | **Argument and Structure** | **Critical Evaluation** | **Presentation** | **Relevance to Literature** |
| Pass  Pass | 86 – 100% | The work examined is exemplary and provides clear evidence of a complete grasp of the knowledge, understanding and skills appropriate to the Level of the qualification. There is also unequivocal evidence showing that all the learning outcomes and responsibilities appropriate to that Level are fully satisfied. At this level it is expected that the work will be exemplary in all the categories cited above. It will demonstrate a particularly compelling evaluation, originality, and elegance of argument, interpretation or discourse. | | | | | | |
| 76 – 85% | The work examined is excellent and demonstrates comprehensive knowledge, understanding and skills appropriate to the Level of the qualification. There is also excellent evidence showing that all the learning outcomes and responsibilities appropriate to that level are fully satisfied. At this level it is expected that the work will be excellent  in the majority of the categories cited above or by demonstrating particularly compelling evaluation and elegance of argument, interpretation or discourse and some evidence of originality | | | | | | |
| 70 – 75% | The work examined is of a high standard and there is evidence of comprehensive knowledge, understanding and skills appropriate to the Level of the qualification. There is  clearly articulated evidence demonstrating that all the learning outcomes and responsibilities appropriate to that level are satisfied At this level it is expected that the standard of the work will be high in the majority of the categories cited above or by demonstrating particularly compelling evaluation and elegance of argument, interpretation or discourse | | | | | | |
| 60 – 69% | Directly relevant to the requirements of the assessment | A substantial knowledge of relevant material, showing a clear grasp of themes, questions and issues therein | Comprehensive analysis - clear and orderly presentation | Well supported, focussed argument which is clear and logically structured. | Contains distinctive or independent thinking; and begins to formulate an independent position in relation to theory and/or practice. | Well written, with standard spelling and grammar, in a readable style with acceptable format | Critical appraisal of up-to-date and/or appropriate literature. Recognition of different perspectives. Very good use of a wide range of sophisticated source material. |
| 50 – 59% | Some attempt to address the requirements of the assessment: may drift away from this in less focused passages | Adequate knowledge of a fair range of relevant material, with intermittent evidence of an appreciation of its significance | Significant analytical treatment which has a clear purpose | Generally coherent and logically structured, using an appropriate mode of argument and/or theoretical mode(s) | May contain some distinctive or independent thinking; may begin to formulate an independent position in relation to theory and/or practice. | Competently written, with only minor lapses from standard grammar, with acceptable format | Uses a good variety of literature which includes recent texts and/or appropriate literature, including a substantive amount beyond library texts. Competent use of source material. |
| 40 – 49% | Some correlation with the requirements of the assessment but there are instances of irrelevance | Basic understanding of the subject but addressing a limited range of material | Some analytical treatment, but may be prone to description, or to narrative, which lacks clear analytical purpose | Some attempt to construct a coherent argument, but may suffer loss of focus and consistency, with issues at stake stated only vaguely, or theoretical mode(s) couched in simplistic terms | Sound work which expresses a coherent position only in broad terms and in uncritical conformity to one or more standard views of the topic | A simple basic style but with significant deficiencies in expression or format that may pose obstacles for the reader | Evidence of use of appropriate literature which goes beyond that referred to by the tutor. Frequently only uses a single source to support a point. |
| Fail | 35 – 39% | Relevance to the requirements of the assessment may be very intermittent, and may be reduced to its vaguest and least challenging terms | A limited understanding of a narrow range of material | Largely descriptive or narrative, with little evidence of analysis | A basic argument is evident, but mainly supported by assertion and there may be a lack of clarity and coherence | Some evidence of a view starting to be formed but mainly derivative. | Numerous deficiencies in expression and presentation; the writer may achieve clarity (if at all) only by using a simplistic or repetitious style | Barely adequate use of literature. Over reliance on material provided by the tutor. |
| 30 – 34% | The work examined provides insufficient evidence of the knowledge, understanding and skills appropriate to the Level of the qualification. The evidence provided shows that some of the learning outcomes and responsibilities appropriate to that Level are satisfied. The work will be weak in some of the indicators. | | | | | | |
| 15 – 29% | The work examined is unacceptable and provides little evidence of the knowledge, understanding and skills appropriate to the Level of the qualification. The evidence shows that few of the learning outcomes and responsibilities appropriate to that Level are satisfied. The work will be weak in several of the indicators. | | | | | | |
| 0 – 14/% | The work examined is unacceptable and provides almost no evidence of the knowledge, understanding and skills appropriate to the Level of the qualification. The evidence fails to show that any of the learning outcomes and responsibilities appropriate to that Level are satisfied. The work will be weak in the majority or all of the indicators. | | | | | | |

**Comments**