Coursework Assignment Brief

Postgraduate

Academic Year 2020-21

Module Title:	Advanced Data Science	
Module Code:	71662	
Assessment Title:		
Assessment Type	CWRK001	Weighting: 100%
School:	School of Computing and Digital Technology	
Module Co-ordinator:	Kuldeep Singha	
Hand in deadline date:	12pm Mid-day on 15th April 2021	
Return of Feedback date and format	20 working days from date of submission (see Moodle for details).	
Re-assessment hand in deadline date:	12pm Mid-day on Monday 22th July 2021 Note: the reassessment work may be different.	
Support available for students required to submit a re-assessment:	Timetabled revisions sessions will be arranged for the period immediately preceding the hand in date	
NOTE:	At the first assessment attempt, the full range of marks is available. At the re-assessment attempt the mark is capped and the maximum mark that can be achieved is 50%.	
Assessment Summary	The coursework involves finding a dataset, formulating a research problem related to this dataset, pre-processing and visualising the data, building and comparing several machine learning models to address the identified problem. The submission must consist of a project report, a dataset and a collection of Python scripts with comments, all included into one zip-file.	

IMPORTANT STATEMENTS

Standard Postgraduate Regulations

Your studies will be governed by CUC Academic Regulations on Assessment, Progression and Awards. Copies of regulations can be found at https://www.coventry.ac.uk/

For courses accredited by professional bodies such as the IET (Institution of Engineering and Technology) there are some derogations from the standard regulations, and these are detailed in your Programme Handbook.

Cheating and Plagiarism

Both cheating and plagiarism are totally unacceptable, and the University maintains a strict policy against them. It is YOUR responsibility to be aware of this policy and to act accordingly. Please refer to the Academic Registry Guidance at https://www.coventry.ac.uk/cuc/legal-documents/academic-and-general-regulations/

The basic principles are:

- Don't pass off anyone else's work as your own, including work from "essay banks". This is plagiarism and is viewed extremely seriously by the University.
- Don't submit a piece of work in whole or in part that has already been submitted for assessment elsewhere. This is called duplication and, like plagiarism, is viewed extremely seriously by the University.
- Always acknowledge all of the sources that you have used in your coursework assignment or project.
- If you are using the exact words of another person, always put them in quotation marks.
- Check that you know whether the coursework is to be produced individually or whether you can work with others.
- If you are doing group work, be sure about what you are supposed to do on your own.
- Never make up or falsify data to prove your point.
- Never allow others to copy your work.
- Never lend disks, memory sticks or copies of your coursework to any other student in the University; this may lead you being accused of collusion.

By submitting coursework, either physically or electronically, you are confirming that it is your own work (or, in the case of a group submission, that it is the result of joint work undertaken by members of the group that you represent) and that you have read and understand the University's guidance on plagiarism and cheating.

You should be aware that coursework may be submitted to an electronic detection system in order to help ascertain if any plagiarised material is present. You may check your own work prior to submission using Turnitin at the https://openmoodle.coventry.ac.uk/. If you have queries about what constitutes plagiarism, please speak to your module tutor or the Centre for Academic Success.

Electronic Submission of Work

It is your responsibility to ensure that work submitted in electronic format can be opened on a faculty computer and to check that any electronic submissions have been successfully uploaded. If it cannot be opened it will not be marked. Any required file formats will be specified in the assignment brief and failure to comply with these submission requirements will result in work not being marked. You must retain a copy of all electronic work you have submitted and re-submit if requested.

Learning Outcomes to be Assessed:

1) Compare the different aspects involved in the modern Data Science.

2) Critically evaluate and practice implementing a wide range of algorithms and modern tools used to solve various data science tasks.

3) Apply learned techniques to formulate and solve real-life data-based problems.

4) Communicate technical information in a range of formats appropriate to a specific audience.

Assessment Details

Title: Building and evaluating data analysis and machine learning pipelines.

Style: Coursework consisting of a report, dataset and programming scripts.

Rationale:

This coursework is most suited for assessing the learning outcomes of the module providing the practical nature of the Data Science field. The area is growing fast and the interest in data analysis and machine learning solutions constantly increases. Learning to formulate and solving practical and research-oriented data-driven projects will ensure your continuing employability through development of analytical soft skills.

Description:

You are required to find a dataset, formulate a problem you want to address with the dataset (e.g. predict whether a mushroom is poisonous or not based on its characteristics), build and evaluate at least two machine learning models that would address the problem, and draw conclusions and recommendations based on your findings. The submission should include your report, dataset (plus any number of sets representing pre-processing stages if needed) and Python scripts with comments, all included in one zip-file. Your work should be original and produced by you. Copying whole tutorials, scripts or images from other sources is not allowed. Any material you borrow from other sources to build on should be clearly referenced (use comments to reference in Python scripts); otherwise, it will be treated as plagiarism, which may lead to investigation and subsequent action.

You can use any open data, e.g.:
https://archive.ics.uci.edu/ml/datasets.php
https://www.kaggle.com/datasets
https://data.gov.uk/

Additional information

Recommended Report Structure:

1. Cover page with title of your project; module code, title, coordinator name; your name and student number; date.
2. Abstract
3. Introduction, background, aim and objectives
4. Dataset(s) description *(can be supported with figures and references to Python code)*
5. Problem to be addressed *(justified and supported with references to literature)*
6. Machine learning model N (*iterate for each model/algorithm)*
    1. Summary of the approach *(justified and supported with references)*
    2. Data pre-processing, visualisation, feature selection *(with references to Python code)*
    3. Model training, evaluation and testing *(with references to Python code)*
    4. Results and discussion *(supported with tables, figures and references to Python code)*
7. Results comparison across the models built (*supported with tables, figures and Python code)*
8. Conclusion, recommendations and future work
9. References

For advice on writing style, referencing and academic skills, please make use of the Centre for Academic Success: https://www.coventry.ac.uk/cuc/legal-documents/academic-and-general-regulations/

Workload:

Recommended length of the report is 3,000 words excluding figures and tables. A typical student would be expected to spend a minimum of 40 hours working on the coursework to pass this assignment.

Transferable skills:

- Problem solving
- Time keeping
- Project management
- Written communication skills

Marking Criteria:

Table of Assessment Criteria and Associated Grading Criteria

Assessment Criteria →	Dataset(s) & Question(s)	Modelling	Code	Report
Weighting	**20%**	**40%**	**20%**	**20%**
Grading Criteria **0 – 29%** **F**	Inappropriate dataset or lack of its initial analysis and understanding; ill-formulated questions.	Missing or inappropriate data pre-processing, feature selection, modelling and/or results interpretation.	Missing or not compiling/executing.	Not appropriately structured with main sections missing.
30 – 39% **E**	Appropriate dataset, but its initial analysis is poor, and/or oversimplified questions.	Incomplete or significant errors in data pre-processing, modelling and/or results interpretation.	Compiling and executing, but implementing only some deliverables.	Badly planned and/or some sections and/or referencing to code missing.
40 – 49% **D**	Satisfactory dataset and questions, but significant errors in initial dataset analysis or not fully justified questions.	Satisfactory data pre-processing, feature selection, modelling and results interpretation, but with some major errors or missing details.	All deliverables are implemented, but there are some major errors, s/w principles are not followed, and/or lack of comments.	All required sections are covered, but structure is not well planned or major details missing.
50 – 59% **C**	Satisfactory dataset and justified questions, but some minor errors in initial analysis.	Good data pre-processing, feature selection, modelling and results interpretation, but with some minor errors or missing details.	All deliverables are implemented, but there are some minor errors, not all s/w principles are followed, and/or insufficient/inaccurate comments.	Well planned with all required sections present, but some details or code referencing missing or not clearly explained.
60 – 69% **B**	Good choice of dataset and questions with fair impact and no errors in initial analysis.	Good data pre-processing, feature selection, modelling and results interpretation, with no errors.	All deliverables are implemented with no errors, but code is not optimised and/or with insufficient comments.	Well planned and clearly formulated with all required sections present, but with some minor details missing.
70 – 79% **A**	Very good choice of dataset and questions with significant impact, no errors in initial analysis.	Very strong case of pre-processing, feature selection, modelling and results interpretation, with attention to detail and no errors.	All deliverables are implemented in efficient way, following s/w principles, with clear and accurate comments, and no errors.	Very well planned and clearly presented, with appropriate and sufficient referencing to code and literature.
80 – 90% **A+**	Excellent choice of dataset and questions with major impact, no errors in initial analysis.	Excellent pre-processing, feature selection, modelling and results interpretation, error-free with some advanced techniques employed and several settings tested.	All deliverables are implemented in efficient way, following s/w principles, employing some advanced methods, with clear and accurate comments, and no errors.	Excellent, complete, clearly presented professional work, with appropriate and sufficient referencing to code and literature.
90 – 100% **A***	Outstanding choice of dataset and questions with significant impact, no errors in initial analysis.	Outstanding pre-processing, feature selection, modelling and results interpretation, error free with some novel techniques employed suitable for publication.	All deliverables are implemented in efficient way, following s/w principles, employing some advanced/novel methods, with clear and accurate comments, and no errors.	Outstanding, complete, clearly presented professional work, with appropriate referencing to code and literature, and suitable for publication.