

CS5811 Distributed Data Analysis

Coursework for 2022/23

TABLE OF CONTENTS

Main Objective of the Assessment.....	1
Description of the Assessment.....	1
Learning Outcomes and Marking Criteria.....	2
Format of the Assessment.....	3
Submission Instructions.....	3
Avoiding Academic Misconduct.....	3
Late Coursework.....	3

Assessment Title	Distributed Data Analysis
Module Leader	Dr. Stasha Lauria
Distribution Date	N/A
Submission Deadline	Monday 24.04.2023 @ 11.00am (week 32)
Feedback by	24.05.2023
Contribution to overall module assessment	100%
Indicative student time working on assessment	Up to 150 Hours
Word or Page Limit (if applicable)	12 Pages (not including references)
Assessment Type (individual or group)	Individual

MAIN OBJECTIVE OF THE ASSESSMENT

The aim of this assignment is to generate value and insight from the processing of heterogeneous data. This will be achieved by implementing several analytic methods/techniques/algorithms, evaluating them and comparing the effectiveness of the adopted approaches.

The development and implementation of the data analysis project will be supported by team-based effort and weekly meetings.

DESCRIPTION OF THE ASSESSMENT

The final report should be an original and individual submission, but it will be underpinned by a group effort and by effective sharing of data and partial results. It is important that *individual contributions shared among the group are clearly defined* (see "Authorship Contribution" below for further details). These contributions should be agreed upfront in a designated meeting.

The management of the data during (and after the project) can be described in a formal *Data Management Plan* (DMP). A template is available on Blackboard Learn. The DMP should be discussed with the group in a designated meeting.

The following parts should be *developed by shared group effort*:

- Data collection
- Data preparation and cleaning
- Exploratory data analysis

Each member of the group is expected to implement and apply at least **two methods/approaches** including:

- One machine learning method for prediction (regression or classification)
- One High Performance Computational technique for distributed data analysis. It is expected that an alternative method/approach to Hadoop will be used

For **both** of these two methods/approaches, each member of the group should produce their own results and distribute/exchange their results among all the other members. Finally, each member is expected to independently compare, discuss and evaluate these shared results.

Submissions will be graded on technical ability, creativity, practicality, and their use of concepts introduced in different study blocks, in particular CS5706 Machine Learning and CS5710 High Performance Computational Infrastructures.



The report should provide an "Authorship Contribution" statement (ACS). This statement should clarify how data generation and/or analyses made by other members of the group has contributed to the project.

AUTHORSHIP CONTRIBUTION: Authorship should meet all 2 of the following conditions:

1. Authors make substantial contributions to conception and design, and/or acquisition of data, and/or analysis and interpretation of data and/or evaluation, and/or visualisation;
2. Authors participate in discussing the results critically for important intellectual content;

Example of "Authorship Contribution" statement (ACS): Y.O. and Y.Z. designed the data collection. G.S., M.K.R. and Y.M. performed the exploratory data analysis. Y.O. implemented and applied the random forest predictor.

LEARNING OUTCOMES AND MARKING CRITERIA

LO1: Design and implement a data analytics solution for generating value and insight from the processing of heterogeneous data using statistical learning and distributed computing technologies.

LO2: Critically evaluate and reflect on the appropriate use of methods and technologies for distributed data analysis, their ability to deliver accurate predictions and the value and limitations of prediction.

The coursework will be marked according to the following criteria:

- a) Identifying a data analytics problem and formulating a relevant research question and plan [LO2]
- b) Preparing, integrating and exploring the data sets suitable to answer the research question [LO1]
- c) Implementing and executing a complete and coherent data analysis [LO1]
- d) Critically reflecting on the results of the data analysis (accuracy, limitations and interpretation) [LO2]

Descriptors for lower grade bands should be satisfied and evidenced for higher grade band award, i.e. B-band grades can only be awarded if all descriptors for C-band and D-band grades have been satisfied and evidenced. Within a grade band, all descriptors should be satisfied and evidenced for the +, three descriptors should be satisfied for the base grade, while the - grade requires at least two out of four of descriptors.

Grade Descriptors	marker discretion to apply +/- grades
The report is incomplete. No or confusing structure in the report. Key aspects of the report (such as problem definition, research question, data preparation/integration, etc) are either missing or confusing. No evidence of implementation. Authorship Contribution is contradictory, confusing or incomplete. One or more of these criteria may apply.	E/F-grade
A clearly written scientific report including all required sections and demonstrating: <ol style="list-style-type: none"> a) correct definition of the problem and formulation of the research question b) basic data preparation and dataset integration c) correct application of one machine learning method and one appropriate HPC technique d) Authorship Contribution statement and reflection on the accuracy of the results 	D-grade (D-, D, D+)
All the requirements for a D-grade plus evidence of: <ol style="list-style-type: none"> a) consistent structure of the whole data analysis driven by the research question b) use of graphical analysis to gain insight on the data sets at exploratory level c) effective use of performance evaluation for methods comparison d) attempt to provide an interpretation of the results and discussing limitations 	C-grade (C-, C, C+)
All the requirements for a C-grade plus evidence of: <ol style="list-style-type: none"> a) clearly presented justification for most of the data analysis steps b) use of at least one unsupervised learning method for exploratory data analysis c) effective use of appropriate HPC techniques in combination with supervised learning methods d) understanding of the results in the context of the research question 	B-grade (B-, B, B+)
All the requirements for a B-grade plus evidence of: <ol style="list-style-type: none"> a) well formulated storytelling about the data across the report and inclusion of DMP 	A-grade (A-, A, A+, A*)



- | | |
|--|--|
| b) use of exploratory data analysis to inform data preparation and/or analysis | |
| c) relevant use of R packages and/or Python libraries | |
| d) new knowledge discovery directly obtained from the data analysis | |

FORMAT OF THE ASSESSMENT

The report should be submitted as a single PDF file. **The report should include exactly the following sections:**

1. *Data description and research question*
2. *Data preparation and cleaning*
3. *Exploratory data analysis*
4. *Machine learning prediction*
5. *High Performance Computational implementation*
6. *Performance evaluation and comparison of methods*
7. *Discussion of the findings*
8. *Data Management Plan and Author Contribution statement*

The main text of the report (including the eight sections above) should not be more than 12 pages (11pt font minimum, the only content allowed beyond the 12th page is an appendix section). Any software produced should be included as code in the Appendix or uploaded as a separate archive file along the PDF file.

SUBMISSION INSTRUCTIONS

You must submit your coursework as a PDF file on Wiseflow by the submission deadline specified above. You can follow the link to Wiseflow through the module's section on Blackboard Learn or login in directly at <https://uk.wiseflow.net/brunel>. The name of your file should follow the normal convention set out in the student handbook, and must therefore include your student ID number (e.g., 0612345.pdf). It can also include the module code (e.g., CS2001_0612345.pdf).

AVOIDING ACADEMIC MISCONDUCT

Before working on and then submitting your coursework, please ensure that you understand the meaning of [plagiarism, collusion](#), and cheating (including [contract cheating](#)) and the seriousness of these offences. Academic misconduct is serious and being found guilty of it results in penalties that can reduce the class of your degree and may lead to you being expelled from the University. Information on what constitutes academic misconduct and the potential consequences for students can be found in [Senate Regulation 6](#).

You may also find it useful to read this [PowerPoint presentation](#) which explains, in plain English, the different kinds of misconduct, how to avoid (even accidentally) committing them, how we detect misconduct, and the common reasons that students give for engaging in such activities.

If you are experiencing difficulties with any part of your studies, remember there is always help available:

- Speak to your personal tutor. If you're not sure who your tutor is, please ask the Taught Programmes Office (TPOcomputerscience@brunel.ac.uk).
- Alternatively, if you prefer to speak to someone outside of the Department you can contact the [Student Support and Welfare](#) team.

LATE COURSEWORK

The clear expectation is that you will submit your coursework by the submission deadline stated in the study guide. In line with the University's policy on the late submission of coursework (revised in July 2016), coursework submitted up to 48 hours late will be accepted, but capped at a threshold pass (D- for undergraduate or C- for postgraduate). Work submitted over 48 hours after the stated deadline will automatically be given a fail grade (F).

Please refer to the [Computer Science student information pages](#) and the [Coursework Submission Procedure](#) pages for information on submitting late work, penalties applied and procedures in the case of Extenuating circumstances.

