University of Glasgow | School of Computing Science

Assessed Coursework

Course Name	**Data Analytics (DA)**			
Coursework Number	**1**			
Deadline	**Time:**	**23:59**	**Date:**	**10th July 2022**
% Contribution to final course mark	**35%**			
Solo or Group	**Solo** ✓		**Group**	
Anticipated Hours	**10 per group member**			
Submission Instructions	**1. Submit via SIT_Xsite a Zip or compressed tar file containing your source code directory (UoG-DA.tgz or UoG-DA.zip), one DAReportSessionxGroupxx.pdf file, Youtube video plus group presentation video(if needed)** **2. Declaration Form**			
Please Note: This Coursework cannot be Re-Done				

Code of Assessment Rules for Coursework Submission

Deadlines for the submission of coursework which is to be formally assessed will be published in course documentation, and work which is submitted later than the deadline will be subject to penalty as set out below.

The primary grade and marks awarded for coursework which is submitted after the published deadline will be calculated as follows:

(i) in respect of work submitted not more than four working days after the deadline

- the work will be assessed in the usual way;
- the primary grade and mark so determined will then be reduced by 15% for each working day (or part of a working day) the work was submitted late.

(ii) work submitted more than four working days after the deadline will be awarded Grade F.

Penalties for late submission of coursework will not be imposed if good cause is established for the late submission. You should submit documents supporting good cause to Admin-In-Charge

Penalty for non-adherence to Submission Instructions is 2 bands
You must complete an "Own Work" form

Data Analytics (UoG-DA) 2021-22

Assessed Exercise: Data Analytics on Predictive Maintenance of Green Energy

Introduction

The goal of this AE exercise is to familiarize yourselves with the design, implementation, and performance testing of the predictive maintenance system of green energy. Given the problem statement with the real dataset, you will be required to go through the whole cycle of problem definition, problem analysis, algorithm, and pseudocode design. You will need to demonstrate the three stages of Data Analytics, namely **D**ata preparation (Data cleaning and preprocessing) , **D**ata Mining and **D**ata Visualization with result analysis.

Assessed task

You will be working in group of 4-5 members according to the respective P Group. Your group are expected to work on the given real-life problem statement with substantial dataset to perform Data Analytics and design a proposed data analytics/machine learning solution. The solution to the problem statement is open ended and not constrained as long as it can be implemented with your own novelty using any data mining algorithms be it classification, probabilistic, regression, association or clustering baselined on data analytics design concept and evaluation.

Every member of your group is expected work cohesively to search, brainstorm to derive the team design that is unique and holistic in solution and implementation. You can search such design online or any research database as following but not limited to. (You can access the IEEE Xplore and ACM using your Glasgow GUID and password).

IEEE Transactions on Big Data

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/search/searchresult.jsp?newsearch=true&queryText=IEEE%20Transactions%20on%20Big%20Data

IEEE Big Data Mining and Analytics

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/xpl/issues?punumber=8254253&isnumber=9430128

IEEE Access

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/xpl/tocresult.jsp?isnumber=9312710&punumber=6287639

ACM SIGKDD International Conference on knowledge discovery and data mining

https://dl-acm-org.ezproxy.lib.gla.ac.uk/conference/kdd/proceedings

IEEE International Conference on Data Mining

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/xpl/conhome/1000179/all-proceedings

ACM Conference on Recommender Systems

https://dl-acm-org.ezproxy.lib.gla.ac.uk/conference/recsys/proceedings

IEEE Conference on Data Science and Advanced Analytics

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/search/searchresult.jsp?newsearch=true&queryText=IEEE%20International%20Conference%20on%20Data%20Science%20and%20Advanced%20Analytics

IEEE Conference on Big Data and Analytics

https://ieeexplore-ieee-org.ezproxy.lib.gla.ac.uk/xpl/conhome/1823704/all-proceedings

In general, the given problem statement must allow the following implementation

1) Data preparation (Data Cleaning and Data Preprocessing)
2) Data Mining Algorithm
3) Data Visualization
4) Data Analysis

Green Energy Preventive Maintenance System

Liquefied Natural Gas (LNG) is identified as one of the green energy sources to reduce global warming. It generates 30% less carbon dioxide than fuel oil and 45% less than coal, with double reduction in nitrogen dioxide emissions and nearly no environmentally damaging sulphur dioxide emission. https://www.youtube.com/watch?v=rjlRTFyennU

The transmission of the LNG requires transmission pipeline that link the regasification

storage to the various destination such as power station, office, factory, and household. The transportation efficiency is monitored by the ultrasonic flowmeters (USM) that are installed along the pipelines. https://www.youtube.com/watch?v=Bx2RnrfLkQg. These USMs are considered the Industrial Internet of Things (IIoT) measuring devices that collect the physical statistics(features/attributes) of the pipelines that transport the LNG fluid and gases. These health statistics (features/attributes) measured how smooth the fluid transmission and can be correlated to the health state of the pipelines. These USM datasets are sent to the central datacentre for health predictive maintenance of entire green energy transportation. In this assessed exercise, there are four USM datasets (TUV-NEL (2012), Testing the diagnostic capabilities of liquid ultrasonic flow meters, National Measurement System) collected for four USM for pipelines namely

Meter A contains **87** instances of physical diagnostic parameters for an **8-path** liquid USM. It has **37** attributes(features) and **2** classes or health states:

Class '1' - Healthy
Class '2' - Installation effects

Meter B contains **92** instances of diagnostic parameters for a **4-path** liquid USM. It has **52** attributes (features) and **3** classes:

Class '1' - Healthy
Class '2' - Gas injection
Class '3' - Waxing

Meter C contains **181** instances of diagnostic parameters for a **4-path** liquid USM. It has **44** attributes (features) and **4** classes:

Class '1' - Healthy
Class '2' - Gas injection
Class '3' - Installation effects
Class '4' - Waxing

Meter D contains **180** instances of diagnostic parameters for a **4-path** liquid USM. It has **44** attributes(features) and **4** classes:

Class '1' - Healthy
Class '2' - Gas injection
Class '3' - Installation effects
Class '4' - Waxing

The attributes(features) of these 4 USM are continuous with the exception of the class attribute.

Meter A
(1) -- Flatness ratio
(2) -- Symmetry
(3) -- Crossflow
(4)-(11) -- Flow velocity in each of the eight paths
(12)-(19) -- Speed of sound in each of the eight paths
(20) -- Average speed of sound in all eight paths
(21)-(36) -- Gain at both ends of each of the eight paths
(37) -- Class attribute or health state of meter: 1,2

Meter B
(1) -- Profile factor
(2) -- Symmetry
(3) -- Crossflow
(4) -- Swirl angle
(5)-(8) -- Flow velocity in each of the four paths
(9) -- Average flow velocity in all four paths
(10)-(13) -- Speed of sound in each of the four paths
(14) -- Average speed of sound in all four paths
(15)-(22) -- Signal strength at both ends of each of the four paths
(23)-(26) -- Turbulence in each of the four paths
(27) -- Meter performance
(28)-(35) -- Signal quality at both ends of each of the four paths
(36)-(43) -- Gain at both ends of each of the four paths
(44)-51 -- Transit time at both ends of each of the four paths
(52) -- Class attribute or health state of meter: 1,2,3

Meters C and D
(1) -- Profile factor
(2) -- Symmetry
(3) -- Crossflow
(4)-(7) -- Flow velocity in each of the four paths
(8)-(11) -- Speed of sound in each of the four paths
(12)-(19) -- Signal strength at both ends of each of the four paths
(20)-(27) -- Signal quality at both ends of each of the four paths
(28)-(35) -- Gain at both ends of each of the four paths
(36)-(43) -- Transit time at both ends of each of the four paths
(44) -- Class attribute or health state of meter: 1,2,3,4

Each group is given **three out of four USM** datasets above and perform supervised (classification or regression) or unsupervised (clustering) learning and create anomaly detection algorithm to ensure timely prediction of anomaly. In terms of implementation, each project group should clearly work out but not limited to

a. Data Preparation/Preprocessing
    I. Decide is there a need for data reduction?
    II. Decide is there a need for data feature extraction?
    III. Show the feature importance (rank the feature importance, namely which feature is the most important and rank accordingly)
    IV. Is there any relationship among the three datasets? Are they coming from the same pipeline or are standalone datasets for different pipelines?
b. Data Mining
    I. Decide whether are you using supervised or unsupervised learning? What algorithm should be preferred for your team?
    II. Decide the split ratio of the training/test dataset such as 70:30 or 80:20?
    III. Determine the classification, anomaly detection performance accuracy such as RMSE, confusion matrix, outlier detection?
c. Data Visualization
    I. Plot the various performance indicators to illustrate the team choice and various

result in part (a) and (b)

d. Result Analysis

I. Justify the results with theoretical understanding.

Group Dataset

P1 Group : Data Set Meter C , Data Set Meter D and Data Set Meter B

P2 Group : Data Set Meter C , Data Set Meter A and Data Set Meter B

P3 Group : Data Set Meter D , Data Set Meter A and Data Set Meter B

What to hand in

According to your Group number, use

- SIT-Xsite Dropbox to submit a single zip or compressed tar file with the contents of the UoG-DA plus a separate DAReportSessionxGroupxx.pdf file and the scanned version of the signed declaration forms of all members in the group

To aid in testing and assessing of your code, please make sure that:

1) Your submission file is named UoG-DA.tgz or UoG-DA.zip.
2) When uncompressed, your files will be in a folder named UoG-DA_SessionxGroupxx. The folder should contain all the following
    a) your dataset
    b) all python source codes that are necessary to take in the dataset and generate all the plots and the printed results.
    c) DAReportSessionxGroupxx.pdf where x and xx indicate your lab session group and group number respectively according to the grouping list.
    d) A group power point presentation video of about 20mins (if situation don't allow physical or online presentation on 7th July that week) where all group members must give their speech on their respective portion of work
    e) A short YouTube video of less than 5mins to demonstrate the problem and the implementation with results.
3) Your DAReportSessionxGroupxx.pdf file should outline and contain
    a) The name of the group member and the **individual contribution**.
    b) your design solution in terms of
        - problem definition,
            - What are your group trying to solve and why is it important to solve this problem?
        - problem analysis,
            - What is the novel solution and its originality that your group is going to propose?
            - How does your group break down the problem and give data analytics approach in providing the solution?
        - algorithm

- What are the algorithms in the solution from data preparation, data mining to the data analysis?
- What are the evaluation criteria such as dissimilarity/similarity measure and test accuracy methodology such as confusion matrix, ROC etc?

  - Pseudocode for the algorithm

c) all the source codes
d) clear mathematical and logical explanation on all the plots and results
e) any interesting aspects of your solution (e.g., assumptions you've made, optimisations that you thought of, etc.),

4) Your submission will be tested for plagiarism. Plagiarism cases will be dealt on a case-by-case basis but suffice to say there will be little tolerance.

How this exercise will be marked

Following timely submission, the exercise will be given a numerical mark between 0 (no submission) and 100 (perfect in every way). The numerical marks will then be converted to a grade. The marking scheme as a group is as follows:

- 10 marks: Approach

  For the approach that clearly aligns and articulates the data analytics concept and methodology, namely the 3D of Data Analytics (Data Preparation, Data Mining and Data Visualization). Marks will be awarded for solution in terms of originality.

- 40 marks: Implementation.

  A clear implementation and articulation on the practice of Data Analytics

  a. Data Preparation
  b. Data Mining
  c. Data Visualization
  d. Result Analysis

- 20 marks: Demonstration and Presentation

  A concise and clear demonstration by the group (and individual) of the proposal and solution in terms

  a. Problem statement and its importance
  b. Proposed solution (with the "3D") and its originality
  c. Proposed Implementation/Methodology and Result Analysis
  d. Demonstration
  e. Conclusion and Future Work

- 30 marks for Report/Code

a. As stated in the report requirement in earlier section on submission on DAReportSessionxGroupxx.pdf file
b. quality and structure of the code; make sure that you use appropriate names for variables/function, library, plot and that your source code is properly structured.

c. comments and documentation in the code and your DAReportSessionGroupxx.pdf file; make sure that you comment and document in your source files, at the very least, the basic steps taken to comment on the function, library and various variables used. Do not make an essay of your code; use your pdf file to discuss further details.

The final marking scheme for individual will be as follows:

Each group member grade = (weighted peer review score by group member with his/her presentation during group presentation) *Group score