University of Lincoln Assessment Framework Assessment Briefing 2021-2022

Module Code & Title: CMP3749M – Big Data

Contribution to Final Module Mark: 100% **10%** for Task 1
90% for Task 2

Description of Assessment Task and Purpose:

You are to submit a single report and associated python source code zip file containing the following two tasks:

Task 1 – Narrative (max 1000 words (**strict)**) (10**%)**

In this task, you should prepare a brief research review in the form of a concise and professional report that considers the four best general approaches (in your opinion) used in big data analysis. To do this, you should identify and select a data science problem scenario from an enterprise or public sector organization. You should then explain each of the four methods you choose in the context of the particular problem scenario you selected. You should consider advantages and disadvantages of each method, a critical reflection including fundamental limitations for the methods you mention.

Please make sure to divide your narrative in several sections and include a table to summarise your discussion. This part of your report document should be clearly identified as "Task One - Narrative" and be delineated from the next task.

Task 2 - Analysis (**strict** max 2500 words) (**90%**)

As a data scientist, your main objective is to organize and analyse data regardless how big or small the data is, often employing typical data science software. The analysis made by a data scientist must be easy enough to understand for all the stakeholders including those who have no knowledge of data science. The objective of this task is to show that you can perform an analysis over a data set to guide the stakeholders to understand the data. **The data can be downloaded from Blackboard in the assessment documents area**. The data needs to be analysed using the data science tools and techniques you were taught in class and contain the information detailed in the Report Guidance (see below).

You are required to write and submit a second section in your report where you need to provide answers to all questions, discuss how you completed the tasks outlined in the report guidance. In addition, you are to provide the Python (Pyspark) code you've developed for these tasks. You are expected to go into sufficient depth to demonstrate knowledge and critical understanding of the relevant processes involved. Note that most of the marks will stem from the clarity of your report, with the source code used as evidence.

You should clearly identify this part of the report document as "Task Two - Analysis" and it should be clearly delineated from the previous task in your final report.

Report Guidance

For the Task 2 element of the report, you should split the discussion into two distinct sections that provide a full and reflective account of the processes undertaken. You are expected to answer all questions in each step in each section in detail, perform all analysis on your own (i.e. as individual work), and provide all Python (PySpark) scripts in one ZIP file as supporting evidence.

The data

It is a dataset of pressurised water reactors (a type of nuclear reactors) with various measurements in different parts of the reactor, including vibration, pressure and power levels. The first column in the spreadsheet indicates the status of the reactor, i.e. 'normal' or 'abnormal'. All the other columns are features which could help us to gain insights into the status of each reactor. You are asked to provide an analysis over this data to discuss if these features could be potentially used to predict whether a reactor is normal or abnormal.

Section I: Data Summary, Understanding and Visualisation (30%)

Download the data set named `nuclear_plants_small_dataset.csv' from Blackboard and save it anywhere on your computer. You need to write Python (PySpark) code to accomplish the following tasks.

As a first step, you need to load the data from the file `nuclear_plants_small_dataset.csv' into a Pyspark DataFrame.

Task 1: Before making any analysis, it is required to know if there are missing values in the data. Are there any missing values? Discuss how you will deal with missing values, even if there are no missing values in this data set.

Task 2: Before making an analysis, it is beneficial to understand the data by looking at the summary statistics. There are two groups of subjects (i.e., the normal group and the abnormal group) in this dataset. For each group, show the following summary statistics for each feature in a table: minimum, maximum, mean, median, mode, and variance values. For each group, plot the box plot for each feature.

Task 3: We want to understand the relationship between features. If two features have high correlations, using only one of them could be enough for our analysis. Show in a table the correlation matrix of the features, where each element in the matrix shows the correlation coefficient of two features. Discuss your observations on the correlation matrix. Are there any features which are highly correlated? In any case, we will use all the features in the following tasks.

Section II: Classification & Big data analysis (60%)

As we have had a preliminary analysis on the data, we want to see if the features could be used to predict the status of a reactor. This is treated as a classification problem.

Task 4: Shuffle the data samples and split it into a 70% training set and a 30% test set. How many examples in each group for the training dataset? How many examples in each group for the testing dataset?

Task 5: Train a decision tree, a support vector machine model and an artificial neural network using the training set, and then apply the trained classifiers to the test set. You will obtain the predicted labels for the test set. Now evaluate the classifiers, respectively, by computing the error rate (`Incorrectly Classified Samples' divided by `Classified Sample'). Calculate the sensitivity and specificity. Discuss the error rate, sensitivity and specificity.

Task 6: Compare the three classifiers based on the results obtained in task 5. Which method would you prefer for classification of this data?

Task 7: Based on the analysis over this data, discuss if these features could be potentially used to detect abnormality in reactors.

Task 8: A larger dataset named `nuclear_plants_big_dataset.csv', which contains more reactor entries, can be downloaded from the Blackboard. Based on this larger dataset, please use MapReduce in Pyspark OR using Hadoop to calculate the minimum, maximum and mean values for every feature.

Learning Outcomes Assessed:

On successful completion of this component a student will have demonstrated competence in the following areas:

- LO1: Critically appraise and evaluate Big Data Analytics concepts, tools and techniques
- LO2: apply data science toolkits in a range of applications and solve real-world problem

Knowledge & Skills Assessed:

Subject Specific Knowledge, Skills and Understanding: Literature searching, Referencing, Numeracy, Project Planning, Techniques and Skills in Data Science, Subject-specific knowledge.

Professional Graduate Skills: Independence and personal responsibility, adaptability, written communication, creativity, critical thinking, IT skills, self-reflection and life-long learning, problem solving, effective time management, working under pressure to meet deadlines.

Emotional Intelligence: Self-awareness, self-management, motivation, resilience, selfconfidence.

Career-focused Skills: Big Data tools, techniques, skills and attributes required by employers, a range of problem strategies to present skills and attributes to employers.

Assessment Submission Instructions:

Report (narrative + analysis)

The submission deadline of this assignment is included in the School Submission dates on Blackboard. You must make an electronic submission of your report including both the narrative and analysis tasks to the Turnitin upload area for Assessment 1.

The report must:

- Contain your name, student number, student email address, and module name;
- Be in single PDF with:
- **no more than 1000 words** (excluding tables and references) for the Task 1 section
- **no more than 2500 words** (excluding tables and references) for the Task 2 section
- Be formatted single-spaced with 11pt font size;

Do not include this briefing document.

Source Code

Your python (Pyspark) code, should be submitted as a single zip archive, to the assessment item 1 supporting documents area on blackboard. This zip archive should contain your python code for all tasks and include code comments where appropriate to aide understanding.

All elements of both tasks are individually assessed. Your work must be presented according to the School of Computer Science guidelines for the presentation of assessed written work. Please make sure you have a clear understanding of the grading principles for this component as detailed in the accompanying Criterion Reference Grid. Your citations and referencing should be in accordance with University guidelines.

If you are unsure about any aspect of this assessment, please seek the advice of the module tutors (contact details on blackboard).

The submission deadline of this assignment is included in the School Submission dates on Blackboard.

If you are unsure about any aspect of this assessment component, please seek the advice of the module lecturers contact details are available on blackboard.

Date for Return of Feedback: Feedback will be provided on blackboard within three weeks of submission (see hand in dates spreadsheet)

Feedback Format:

Summative feedback for Task 1 (narrative) and Task 2 (analysis) will be provided on BlackBoard according to CRG criteria (see CRG file). You will be given formative verbal feedback during the workshop sessions.

Additional Information for Completion of Assessment:

Students are encouraged to use any lecture and their own personal notes to assist them with the completion of the assessment. Also, students are allowed to use any library and/or online resource as a guide on how to solve the assessment problems.

Assessment Support Information:

Students are encouraged to seek assistance from any member of the delivery team and particularly from the module coordinator as means to complete the assessment.

Important Information on Dishonesty & Plagiarism:

University of Lincoln Regulations define plagiarism as 'the passing off of another person's thoughts, ideas, writings or images as one's own...Examples of plagiarism include the unacknowledged use of another person's material whether in original or summary form. Plagiarism also includes the copying of another student's work'.

Plagiarism is a serious offence and is treated by the University as a form of academic dishonesty. Students are directed to the University Regulations for details of the procedures and penalties involved.

For further information, see www.plagiarism.org