Coursework Specification

Read this coursework specification carefully, it tells you how you are going to be assessed, how to submit your coursework on-time and how (and when) you’ll receive your marks and feedback.

|  |  |
| --- | --- |
| **Module Code** | CSI\_6\_DMA |
| **Module Title** | Data Mining and Big Data Analytics |
| **Lecturer** |  |
| **% of Module Mark** | 60% |
| **Distributed** | 11/10/2024 |
| **Submission Method** | Submit online via this Module’s Moodle site |
| **Submission Deadline** | 17:00, 12/12/2024 |
| **Release of Feedback & Marks** | Feedback and provisional marks will be available in  **Grades** on Moodle from 06/01/2025 |

# Coursework Aim:

This individual coursework project is about using analytics to address real-world problems for insight creation. The aim of this coursework is to evaluate your understanding of the basic theories, concepts, methodologies, and the typical algorithms in data mining, and your skills of using Python for analytics.

# Coursework Details:

|  |  |
| --- | --- |
| **Type:** | Project report |
| **Overall View:** | This assignment is to be undertaken **individually**. You should plan your work carefully following the module’s weekly teaching programme and have a regular discussion with your tutor to address any questions and issues you may have during this project. You are expected to produce a written report for this data mining project.  The assignment involves analysing a real-world dataset to identify its underlying patterns and trends by using appropriate data mining techniques and algorithms. These |

|  |  |
| --- | --- |
|  | patterns and trends are intended to be used to address certain business concerns. A dataset will be assigned to you by your tutor. |
| **Tasks:** | You are required to undertake the following tasks in this project:   1. **Business Understanding**    * Download the dataset assigned to you from the module Moodle site along with the data description file (i.e., the metadata).    * Read the data description to learn the nature of the dataset, such as what it is about, where it comes from, and what certain business context it is associated with, etc.    * Examine the dataset within its business context and identify certain meaningful problems that potentially can be addressed by analysing the dataset.    * Translate the business problems to appropriate data mining problems.    * You may also refer to any articles in the literature related to the dataset under consideration. 2. **Data Understanding**    * Perform initial data exploration to get to know more about the dataset, such as the total number of instances in the dataset, the number of attributes (variables), and the data type of each attribute.    * Identify any data quality issues within the dataset, including incorrect data types, missing values, outliers, extreme values, incomparable value ranges of variables, and imbalanced classes, etc. **Note**: Make correction to any incorrect data types before proceeding any further.    * Explore the basic statistics of each attribute, including value range, average, standard deviation, skewness, kurtosis, and mode, etc.    * Determine if the dataset is appropriate to be used for addressing the business problems you have identified in Task 1. If not, re-do Task 1. 3. **Data Preparation**    * Choose appropriate methods for data pre- processing, which includes changing data types, dealing with missing values, tackling outliers, extreme values, and imbalanced classes, conducting data transformation and normalisation, and reducing dimensionality, etc., wherever appropriate.    * Identify correlations among certain variables. |

|  |  |
| --- | --- |
|  | * Determine which, why, and how each attribute should (not) be used in your analysis. * Divide the whole dataset into several subsets to be used for training, test and validation in predictive modelling.  1. **Modelling**    * Use the pre-processed dataset to perform the data mining tasks you have identified in Task 1.    * Choose appropriate techniques and algorithms for your analysis: Choose either *k*-means clustering ***OR*** association rule analysis for descriptive modelling; Choose either decision tree ***OR*** regression for predictive modelling.    * Determine appropriate settings of the algorithms to be applied, e.g., how many clusters to use in *k*- means clustering.    * Re-do data preparation in Task 3 if needed. 2. **Evaluation**    * Provide an explicit and concise description and explanation of ***Both*** the descriptive and predictive models you have created. Examine and explain what patterns and insight you have identified.    * Evaluate the performance of the predictive models in terms of evaluate the performance of the predictive models in terms of various measures applicable, such as accuracy, SSE (sum of squared errors), generalisation ability, simplicity and cost etc.    * Discuss how the descriptive and predictive models created can be used to address the original business problems identified in Task 1.    * Summarise your main findings from the project. |
| **Word Count:** | As a guide, aim for 2500-3000 words. The maximum word limit is 3000 words. If the total word limit is exceeded, it will affect the marks awarded to the project presentation.  **In your report, do not include and explain any basic concepts of the subject; Instead, you should demonstrate your understanding of the concepts by applying them appropriately in your coursework.**  Footnotes will not count towards word count totals but must only be used for referencing, not for the provision of additional text. The bibliography will not count towards the word total. |
| **Presentation:** | * Report must contain Title page, Table of Contents, Abstract, Conclusion, and References. |

|  |  |
| --- | --- |
|  | * Work must be referenced, and a bibliography provided. * Work must be submitted as a Word document (.doc/docx) or a PDF. * Course work must be submitted using Arial font size 11 (or larger if you need to), with a minimum of 1.5 line spacing. * Your student number must appear at the front of the coursework. Your name must **not** be on your coursework. |
| **Referencing:** | Harvard Referencing should be used, see your [Library](https://libguides.lsbu.ac.uk/subjects/home) [Subject Guide](https://libguides.lsbu.ac.uk/subjects/home) for guides and tips on referencing. |
| **Regulations:** | Make sure you understand and expected academic practice and academic misconduct. Note in particular:   * Your work must be your own. Markers will be attentive to both the plausibility of the sources provided as well as the consistency and approach to writing of the work. Simply, if you do the research and reading, and then write it up on your own, giving the reference to sources, you will approach the work in the appropriate way and will cause not give markers reason to question the authenticity of the work. * All quotations must be credited and properly referenced. Paraphrasing is still regarded as plagiarism if you fail to acknowledge the source for the ideas being expressed.   **TURNITIN:** When you upload your work to the Moodle site it will be checked by anti-plagiarism software. |

Learning Outcomes

This coursework will partially assess the following learning outcomes for this module as indicated by **\***.

## Knowledge and Understanding

On successful completion of this module, you will be able to:

* Describe and explain the concepts of data mining including the techniques and algorithms for problem solving and creating competitive advantage. **\***

## Intellectual Skills

On successful completion of this module, you will be able to:

* Critically evaluate different types of data mining tasks in relation to various business and scientific problems, including descriptive modelling and predictive modelling, including cluster analysis, association analysis, and decision and regression for classification and prediction. **\***

## Practical Skills

On successful completion of this module, you will be able to:

* Transfer a business and/or scientific problem into an appropriate data mining problem. **\***
* Creatively apply data mining tools and platforms such as Python package.**\***

## Transferable Skills

On successful completion of this module, you will be able to:

* Analyse and develop solutions for a wide range of business and scientific problems. **\***

# Assessment Criteria and Weighting

University marking criteria have been developed to help tutors give you clear and helpful feedback on your work. They will be applied to your work to help you understand what you have accomplished, how any mark given was arrived at, and how you can improve your work in future.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Criteria** | **Feedforward comments** |  | | | | | |
| **100 - 80%** | **79 - 70%** | **69 - 60%** | **59 - 50%** | **49 - 40%** | **39 - 30%** | **29 - 0%** |
| **10%** | **1. Business** | Exceptionally thorough and clear | Thorough and clear analysis | Clear analysis of business | Clear analysis of | Basic analysis of the key | Inadequate analysis of | Little or no analysis |
| **Understanding** | analysis of business concerns | of business concerns and | concerns and associated | business concerns and | business concerns and | business concerns and | of business |
|  | and associated data mining | associated data mining | data mining tasks to a | associated data mining | associated data mining | associated data mining | concerns and |
|  | tasks. | tasks. | certain depth. | tasks. Probably lack | tasks. | tasks. Lack clarity and | associated data |
|  |  |  |  | some in-depth view. |  | relevance. | mining tasks. |
| **10%** | **2. Data Understanding** | Exceptionally excellent and | Excellent initial data | Good initial data exploration | Essential initial data | Limited simple initial data | Inadequate and/or | Little or no initial |
|  | creative initial data exploration | exploration with effective | performed with appropriate | exploration performed. | exploration. Probably lack | inappropriate initial | data exploration |
|  | with effective means. Thorough | means. Thorough summary | means. Clear summary of | Essential analysis of | some relevance and/or | data exploration | performed. Little or |
|  | summary of the dataset. | of the dataset. Excellent | the dataset. Good analysis | data quality issues and | clarity. Limited use of | performed. Lack clarity | no relevancy. No |
|  | Excellent analysis of data quality | analysis of data quality | of data quality issues and | the role of each | Python. | and relevance. Poor | or inappropriate |
|  | issues and the role of each | issues and the role of each | the role of each attribute. | attribute. Good use of |  | use of Python. | use of Python. |
|  | attribute. Excellent use of | attribute. Excellent use of | Good and flexible use of | Python. |  |  |  |
|  | Python. | Python. | Python. |  |  |  |  |
| **25%** | **3. Data Pre-processing** | Exceptionally thorough and | Thorough consideration of | Good consideration of data | Reasonable | Limited consideration of | Inadequate and/or | Little or no data |
|  | extensive consideration of data | data quality issues for pre- | quality issues for pre- | consideration of data | data quality issues for | inappropriate view of | quality issues |
|  | quality issues for pre-processing. | processing. Appropriate | processing. Appropriate | quality issues for pre- | pre-processing. Some | data quality issues. | considered. |
|  | Appropriate approaches adopted | approaches adopted with | approaches adopted with | processing. Appropriate | appropriate approaches | Inappropriate | Inappropriate |
|  | with exceptionally clear | outstanding understanding. | clear understanding and | approaches adopted | adopted with limited | approaches adopted. | approaches |
|  | understanding. Excellent use of | Excellent use of Python. | every aspect covered. Good | with reasonable | understanding and limited | Poor use of Python. | adopted. No or |
|  | Python. |  | and flexible use of Python. | understanding and | coverage. Limited use of |  | inappropriate use |
|  |  |  |  | most of the main issues | Python. |  | of Python. |
|  |  |  |  | covered. Good use of |  |  |  |
|  |  |  |  | Python. |  |  |  |
| **20%** | **4. Modelling** | Appropriate algorithms | Appropriate algorithms | Appropriate algorithms | Appropriate algorithms | Some appropriate | Inappropriate and/or | Little or no |
|  | employed with exceptionally | employed with clear | employed with clear | employed with | algorithms employed with | inadequate algorithms | algorithms |
|  | clear understanding. Modelling | understanding. Modelling | understanding. Good and | reasonable | limited understanding. | employed. Poor use of | employed. Little or |
|  | with excellent working | with excellent working | flexible use of Python. | understanding. Good | Limited use of Python. | Python. | no use of Python. |
|  | knowledge of Python | knowledge of SAS |  | use of Python. |  |  |  |
|  |  | Enterprise Miner. |  |  |  |  |  |
| **15%** | **5. Model Evaluation** | Exceptionally thorough and clear | Thorough and clear model | Clear model interpretation | Basic model | Weak model | Poor model | Little or no model |
|  | model interpretation and | interpretation and | and comparison with | interpretation and | interpretation and | interpretation and | interpretation and |
|  | comparison with regards to | comparison with regards to | regards to business | comparison with | comparison with regards | comparison with | comparison with |
|  | business concerns. Excellent | business concerns. | concerns. Significantly | regards to business | to business concerns. | regards to business | regards to |
|  | and meaningful models/patterns | Excellent meaningful | meaningful models/patterns | concerns. Reasonable | Very limited | concerns. No or little | business concerns. |
|  | created. | models/patterns created. | created. | models/patterns | meaningfulness. Probably | meaningful |  |
|  |  |  |  | created. | lack some clarity. | models/patterns |  |
|  |  |  |  |  |  | provided. |  |
| **20%** | **6. Report** | Exceptionally clear and concise | Very clear and concise | Clear and concise summary | Clear review and | Adequate review of | Inadequate review of | Little or no review |
|  | summary of project findings. | summary of project findings. | of project findings. Excellent | summary of project | project findings. Probably | project findings. Lack of | of project findings. |
|  | May raise questions for future | May raise questions for | presentation. Clear structure | findings. Good | lack of some clarity. | clarity and accuracy. | Significantly Lack |
|  | research. Exceptional | future research. | and layout. | presentation with | Acceptable presentation. | Poor presentation. | of clarity and |
|  | outstanding presentation. Clear | Outstanding presentation. |  | proper structure and |  |  | accuracy. Very |
|  | structure and layout. | Clear structure and layout. |  | layout. |  |  | poor presentation. |

# How to get help

We will discuss this Coursework Specification in class. However, if you have related questions, please contact me [name and email] as soon as possible.

# Resources

List resources such as background reading, templates, samples, tools, videos, links, etc

# Quality assurance of coursework specifications

Coursework specifications within CSI division go through internal (for new modules with 100% coursework also through external) moderation. This is to ensure high quality, consistency and appropriateness of the coursework as well as to share best practice within the CSI division.