**A black text on a white background

Description automatically generated**

**DEGREE PROGRAMMES**

**GROUP ASSIGNMENT**

**FEBRUARY 2025**

**MODULE NAME : DATA ANALYTICS & MACHINE LEARNING**

**MODULE CODE : ITS69304**

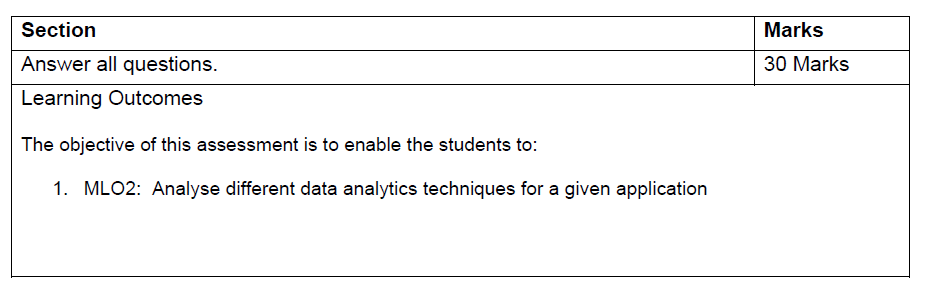
**DEADLINE : /02/2025 8:00AM (MYT GMT+8)**

**This paper consists of SIX (6) printed pages, inclusive of this page.**

**Instruction to Candidates:**

1. Name your answer file as ITS69304\_XXXXXX\_GROUPASGNMT.pdf where XXXXXX is your GROUP NUMBER. Then, submit to the MyTIMeS portal via the link “GROUP ASSIGNMENT submission” on the module page. (Do not submit the question paper).
2. This is a group assignment consisting of 3-5 members per group. Please assign one group leader and put all your group member’s names in the link shared in MyTimes.
3. Each group needs to submit (1) Answer script in pdf format (2) Video of presentation explaining each question.
4. Please write all group members name with the student declaration as per below on your cover page of your answer script.



****

**Sentiment Analysis for Mental Health**

Mental health is more than the absence of mental disorders. It exists on a complex continuum, which is experienced differently from one person to the next, with varying degrees of difficulty and distress and potentially very different social and clinical outcomes. Mental health conditions include mental disorders and psychosocial disabilities as well as other mental states associated with significant distress, impairment in functioning, or risk of self-harm. People with mental health conditions are more likely to experience lower levels of mental well-being, but this is not always or necessarily the case.

Determinants of mental health

Throughout our lives, multiple individual, social and structural determinants may combine to protect or undermine our mental health and shift our position on the mental health continuum. Individual psychological and biological factors such as emotional skills, substance use and genetics can make people more vulnerable to mental health problems. Exposure to unfavourable social, economic, geopolitical and environmental circumstances – including poverty, violence, inequality and environmental deprivation – also increases people’s risk of experiencing mental health conditions. Risks can manifest themselves at all stages of life, but those that occur during developmentally sensitive periods, especially early childhood, are particularly detrimental. For example, harsh parenting and physical punishment is known to undermine child health and bullying is a leading risk factor for mental health conditions. Protective factors similarly occur throughout our lives and serve to strengthen resilience. They include our individual social and emotional skills and attributes as well as positive social interactions, quality education, decent work, safe neighbourhoods and community cohesion, among others.

Mental health risks and protective factors can be found in society at different scales. Local threats heighten risk for individuals, families and communities. Global threats heighten risk for whole populations and include economic downturns, disease outbreaks, humanitarian emergencies and forced displacement and the growing climate crisis. Each single risk and protective factor has only limited predictive strength. Most people do not develop a mental health condition despite exposure to a risk factor and many people with no known risk factor still develop a mental health condition. Nonetheless, the interacting determinants of mental health serve to enhance or undermine mental health.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Question 1:*   1. Import the dataset and explore the data from the mentalhealth.csv dataset. Indicate missing values if there are any. From which columns? Calculate how many of them. Clean the dataset from missing values by impute the datum instead of removing them and attach the clean dataset to the final report [6 marks] 2. Status attributes have several categories. Identify the categories and count how many of them are in each category. Rank your answer ascendingly. Make the status into binary classification only (normal and depression). Explain your method. [4 marks]   *Question 2:*   1. Status attribute is in categorical type. Convert the attribute to numerical values. [2 marks] 2. Split dataset to 80% training and 20% testing data. What theory you are using for splitting the dataset. Status is the dependent variable, and statement is the independent variable. Create a model using Decision Tree and Neural Network classification algorithms for this sentiment analysis problem. [8 marks]   *Question 3:*   1. Evaluate the performance of both models using accuracy, precision and recall. [6 marks] 2. Use ROC to evaluate the classification of whether normal or depressing? Plot the graph for ROC. Give your opinion. [4 marks]   *Question 4:*   1. Examine these new test statements below. Show how it has been classified whether normal or depressed using both models (Decision Tree and Neural Network). [10 marks]  * I can't stop worrying about everything * I've been working hard, and seeing the results makes me feel incredibly happy and fulfilled * Even the smallest things feel like too much right now * I can’t stop smiling * Today has been amazing!  |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Marking Rubric**  **ITS69304-DAML**  **GROUP ASSIGNMENT**  **(FEBRUARY 2025)** | | | | | |  | | | | | | **Criteria** | **Score (Percentage of the allocated marks for each task)** | | | | | **Excellent** | **Good** | **Average** | **Poor** | | >= 10% | < 15%  >= 10% | < 10%  >= 5% | < 5% | | Q1: Data Exploration and Data Pre-Processing | All the presented statistics and types of data and dataset(s) are identified correctly. All types of noises or errors are found in those data. All the preprocessing techniques are identified correctly and explained clearly The answer is supported with some samples/records from the dataset(s). The writing is clear and understandable. The similarity is less than 2%. | Most of the presented statistics and types of data and dataset(s) are identified correctly.  Most of the types of noises or errors are found in those data. All the preprocessing techniques are identified correctly and explained clearly The answer is supported with some samples/records from the dataset(s). The writing is clear and understandable. The similarity is less than 4%. | Some of the presented statistics and types of data and dataset(s) are identified correctly.  Some types of noises or errors are found in those data. All the preprocessing techniques are identified correctly and explained clearly The answer is not supported with samples/records from the dataset(s). The writing is understandable. The similarity is less than 5%. | The presented statistics and the types of data and dataset(s) are not identified correctly.  AND/OR some types of noises or errors of the dataset are not highlighted. All the preprocessing techniques are identified correctly and explained clearly The answer is not supported with samples/records from the dataset(s). The writing is not clear and understandable. The  similarity is more than 5%. | | Q2: Data modelling and development | The data modelling is valid and acceptable. The program/code is correct. The justification is accurate, clear, and valid. The similarity is  less than 2%. | The data modelling is valid and acceptable. The program/code is correct. The justification is acceptable but not well written. The similarity is  less than 4%. | The data modelling is valid and acceptable. The program/code is correct. But the justification is not well written. The similarity is less than 5%. | The data modelling is not valid and acceptable. The program/code is not correct. The justification is not accurate, not clear, and not valid.  The similarity is more than 5%. | | Q3: Evaluation and Visualizing the performance of the model | The evaluation technique and the visualization are valid and acceptable. The program/code is correct. The justification is accurate, clear, and valid. The similarity is  less than 2%. | The evaluation technique and the visualization are valid and acceptable. The program/code is correct. The justification is acceptable but not well written. The similarity is  less than 4%. | The evaluation technique and the visualization are valid and acceptable. The program/code is correct. But the justification is not well written. The similarity is less than 5%. | The evaluation technique and the visualization are not valid and acceptable. The program/code is not correct. The justification is not accurate, not clear, and not valid.  The similarity is more than 5%. | | Q4: Testing new dataset | The answer to the new testing data is correct. All the provided evidences are relevant and correct. The writing is clear and understandable.  The similarity is less than 2%. | The answer to the new testing data is correct. Most of the provided evidences are relevant and correct the writing is clear and understandable. The  similarity is less than 4%. | The answer to the new testing data is correct. Some of the provided evidences are relevant and correct the writing is clear and understandable. The  similarity is less than 5%. | The answer to the new testing data is correct. A few of the provided evidences are relevant and correct the writing is not clear and understandable. The  similarity is more than 5%. |   **Submission Requirements**   1. Font type : Times New Roman 2. Font size : 12 3. Line spacing : 1.5 4. Alignment : Justify Text 5. Document type : .pdf, .ipynb, video (any type) 6. Number of pages : 5 – 20 pages (do not exceed the page limit) 7. Your full report should consist of the following:    * 1. Cover page (Name, ID, Date, Signature, Score)      2. Marking Rubrics & Declaration (attach as second page in the report)      3. Report of your answer script      4. Appendixes (line spacing = 1.0)      + List of references (APA format)      + Python script      + Report of similarity score (percentage of similarity score from each source needs to be shown) 8. Start each question on a separate page. 9. All figures and tables are labelled properly. 10. File naming conventions: StudentName\_FinalExam   **Notes:**   * Student is not allowed to transcribe directly (copy and paste) any material from another source into their submission. * Start each question on a new page. * Answer in form of short essay (50 to 200 words) and print out the relevant Python program outputs * All process/functions must be clearly explained. * Include in-text citation to support your answers and add the list of references at the end of your report (APA format). The list of references is to be alphabetized by the first author's last name, or (if no author is listed) the organization or title. * The Turnitin similarity for this module is 20% overall and lesser than 1% from a single source excluding program source codes.   **-END OF QUESTION PAPER-** |
|  |