TAYLOR'S UNIVERSITY

Wisdom · Integrity · Excellence

ITS65704 DATA SCIENCE PRINCIPLES

GROUP ASSIGNMENT

HAND OUT DATE: MONDAY, 29 OCTOBER 2024

HAND IN DATE: TUESDAY, 26 NOVEMBER, 5:00PM

Instructions to students:

- The group assignment should be attempted in group of 3-4 members.
- Complete this cover sheet and attach it to your submission – this should be your first page.

Student declaration:	
I declare that: ▪ ***I understand what is meant by plagiarism.*** ▪ ***The implication of plagiarism and usage of AI generative tool have been explained to us by our lecturer.*** ***This project is all our work and I have acknowledged any use of the published or unpublished works of other people.***	
NAME	
Name	**Student ID**

Avoid copy and paste job in your report and it is considered as plagiarism. Plagiarism in all forms is forbidden. Students who submit plagiarised document will deserve 0 marks.

1.0 Objective/Learning Outcomes:

Module Learning Outcome:
MLO2 - *Propose suitable data science-related algorithm(s) or model(s) to solve a problem for a given dataset selected from a specific domain.*

2.0 Information and Submission Instructions:

- Total marks: 100
- Weightage: 30%
- The submission should include the following parts in one PDF file:
  - Part 1: TU's template cover page
  - Part 2: Answers to the questions
- Instructions for Answers/Codes:
  - Ensure that explanations provided are directly related to the problem given in the document
  - Include detail explanations whenever necessary.
- Adherence to Instructions:
  - Follow all instructions given carefully to avoid any mark deductions.
  - Double-check your submission to ensure that all required parts are included and formatted correctly according to the provided instructions.
- Further guideline
  - The submission should only be done by the group leader.
  - As per the School of Computer Science policy, late submission within 12 hours after the actual submission time will have mark deduction of 50%. Any submission beyond 12 hours would not be accepted.
  - Plagiarism involves using others' work without credit.
  - Prioritize originality and integrity.
  - Avoid AI generative techniques to prevent plagiarism.
  - Follow guidelines strictly.
  - Understand consequences for violations.

3.0 Questions:

1.0 Project Overview and Instructions

You are required to prepare a Data Science Principles Project report and Python program demonstrating your skills in data science using Google Colab.

2.0 Project Report Requirements

Your project report should cover the following components, with a total of 100 marks distributed across various sections. Each component is weighted according to its importance, as detailed in the rubric. The report organisation should follow th given rubric.

You may refer to the following resources for project ideas and datasets. However, you can choose data from any suitable source.

Suggested Data Sources:

- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.html
- Kaggle Datasets: https://www.kaggle.com/

Example Datasets:

1. Deep Learning CNN for Brain Tumor Detection: https://www.kaggle.com/code/onesinustamba/deep-learning-cnn-dcnn-for-brain-tumor-detection

2. Medical Image Dataset for Brain Tumor Detection: https://www.kaggle.com/datasets/pkdarabi/medical-image-dataset-brain-tumor-detection/code

3. Melanoma Analysis and Model: https://www.kaggle.com/code/saife245/melanoma-detail-analysis-eda-ip-augmentation-model

4. Smoker Lung Cancer Stage Detection: https://www.kaggle.com/code/sasakitetsuya/smoker-lung-cancer-stage-detection-model

5. COVID mRNA Vaccine Analysis: https://www.kaggle.com/code/itsuki9180/mvan-covid-mrna-vaccine-analysis-notebook-268

6. Pulmonary DICOM Preprocessing: https://www.kaggle.com/code/allunia/pulmonary-dicom-preprocessing

Note:
If using image datasets, consider converting them to CSV files using libraries that support such transformations.

- END OF ASSIGNMENT QUESTIONS –

Criteria	Excellent	Good	Satisfactory	Needs Improvement	Poor
2.1 Background and Project/Business Goal	Provides a clear, detailed background with a well-defined project goal.	Background and goal are clear but lack some detail.	Background and goal are present but vague.	Background or goal is unclear or incomplete.	Background and goal are missing or irrelevant.
Project Background (5%)	Provides comprehensive background information.	Provides sufficient background information.	Background information is basic and lacks depth.	Background information is unclear or incomplete.	Background information is not provided or irrelevant.
Explanation of Project Goal (5%)	Clearly defines and explains the project goal.	Adequately defines the project goal.	Project goal is defined but lacks clarity.	Project goal is vague or incomplete.	Project goal is not defined or relevant.
2.2 Data Set Description	Provides a thorough description of the dataset, including clear examples.	Provides an adequate description of the dataset.	Dataset description is present but lacks depth.	Dataset description is vague or incomplete.	Dataset description is not provided or unclear.
Data Characteristics and Source (5%)	Clearly explains data characteristics and provides a reliable source.	Explains data characteristics but lacks some detail.	Provides basic information on data characteristics.	Data characteristics explanation is vague or unclear.	Data characteristics are not explained or the source is unreliable.
High-Level Statistics (10%)	Presents comprehensive and relevant statistical summaries.	Provides relevant statistical summaries.	Statistical summaries are present but limited.	Statistical summaries are unclear or incomplete.	Statistical summaries are missing or irrelevant.
2.3 Data Preprocessing and Issues	Thoroughly describes data issues and presents well-executed preprocessing steps.	Adequately describes data issues and preprocessing steps.	Data issues and preprocessing steps are present but basic.	Data issues or preprocessing steps are vague or incomplete.	Data issues and preprocessing steps are not provided.
Identification of Data Issues (5%)	Clearly identifies and explains relevant data issues.	Adequately identifies and explains data issues.	Identifies data issues but lacks detail.	Data issues identification is vague or unclear.	Data issues are not identified or explained.

Preprocessing Techniques (10%)	Provides a comprehensive explanation of preprocessing techniques.	Provides sufficient explanation of preprocessing techniques.	Explanation of preprocessing techniques is basic.	Preprocessing techniques explanation is unclear or incomplete.	Preprocessing techniques are not explained or relevant.
2.4 Data Science Techniques	Explains the techniques used with detailed rationale for each choice.	Explains the techniques used with an adequate rationale.	Techniques explanation is present but lacks detail.	Explanation or rationale of techniques is unclear or incomplete.	Techniques explanation is not provided or relevant.
Description of Techniques (20%)	Thoroughly describes the techniques used.	Adequately describes the techniques used.	Techniques description is basic.	Techniques description is vague or unclear.	Techniques description is missing.
Rationale for Technique Selection (15%)	Provides a comprehensive rationale for technique selection.	Provides an adequate rationale for technique selection.	Rationale is present but lacks detail.	Rationale is vague or unclear.	Rationale is not provided.
2.5 Model Validation	Clearly explains the model validation process with relevant methods.	Adequately explains the model validation process.	Explanation of model validation is present but basic.	Explanation is unclear or incomplete.	Model validation process is not explained.
2.6 Conclusion	Provides insightful observations and comprehensive suggestions.	Provides adequate observations and suggestions.	Observations and suggestions are basic.	Observations or suggestions are vague or incomplete.	Observations and suggestions are missing or irrelevant.
Google Colab	Attach your colab file for for review. Make sure the file is not having a restricted access.				