TUNKU ABDUL RAHMAN UNIVERSITY OF MANAGEMENT AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATION TECHNOLOGY

ACADEMIC YEAR 2025/2026

BMDS2003 DATA SCIENCE

COURSE NAME	**:**	**Data Science**
COURSE CODE	**:**	**BMDS2003**
SESSION	**:**	**202509**
TOTAL MARK	**:**	**100%**
WEIGHTAGE TO FINAL MARK	**:**	**42%**
Group	**:**	**3 to 4 members**
SUBMISSION DEADLINE	**:**	**Week 6 (Saturday, 11:59 pm)**
PRESENTATION	**:**	**Week 7**

Learning Outcome Being Assessed:

CLO1:	Apply appropriate data science concepts, statistics, and machine learning to make predictions on data (C3, PLO2).
CLO2:	Practice exploratory data analysis and visualisation methods with the aid of appropriate data science tools (P3, PLO3).
CLO3:	Analyse relevant techniques, statistical methods, and Machine Learning algorithms for data analytics (C4, PLO7).

Overview

This assignment requires students to apply the Cross Industry Standard Process for Data Mining (CRISP-DM) framework to a selected dataset using Python. The goal is to identify a real-world classification or regression problem, apply suitable data science techniques, and produce insights to support business decision-making. Students will work in groups of **THREE (3) to FOUR (4)** to complete the project within the 7-week semester. The project must demonstrate the ability to manage data, build and evaluate machine learning models, and communicate analytical findings effectively.

Group Assignment Instructions

Choose **ONE (1)** dataset from the list provided for your group assignment. Click the link to download the dataset. The following is a list of datasets:

1. Heart Disease
2. Taiwanese Bankruptcy
3. Drug Consumption
4. Obesity Levels
5. Online Gaming
6. Tezpur University Android Malware
7. Communities and Crime
8. Traffic Flow
9. Aids Clinical Trials
10. Productivity of Garment Employees
11. Diamonds
12. Seoul Bike Sharing Demand

Important:

- Each dataset is only allowed to be selected by one group per practical session. **FIRST COME, FIRST SERVED.**
- You may refer to the Google Spreadsheet provided by your tutor to check dataset availability and register your selection.
- This assignment should cover a wide range of content discussed in the lectures and practical sessions.

Assignment Guidelines

A. Written Report

Submit a professionally written report in Google Docs format, structured as follows:

1. Cover Page
   - Project title, course code, group number, and member names and IDs.

2. Executive Summary (½ page)
   - Brief overview of the project, dataset, and main findings.

3. Business Understanding
   - Define the business or analytical problem and objectives.
   - Describe the significance and potential business impact.

4. Data Understanding
   - Describe the dataset source, type, and main features.
   - Load and integrate data from multiple sources where necessary.
   - Include summary statistics and data descriptions.

5. Data Preparation
   - Clean and transform raw data to prepare it for analysis.
   - Address data quality issues, including handling missing values and outliers.
   - Standardise data formats and enrich source data where necessary.
   - Document the preprocessing steps in the report.

6. Modelling
   - Select and justify at least **THREE (3)** machine learning models, one of which should serve as the baseline.
   - Comparisons within the same model family (e.g., Decision Tree, Random Forest, Gradient Boosting) or between different families (e.g., KNN vs Naïve Bayes) are acceptable.
   - Explain parameter tuning and model configurations.
   - Justify model choice with supporting references (at least two research or technical sources).

7. Evaluation
   - Present appropriate performance metrics (e.g., accuracy, precision, recall, RMSE).
   - Compare model results and discuss findings.
   - Reflect on limitations and potential improvements.

8. Conclusion
   - Summarise the key findings of the project, emphasising the advantages and limitations of the models used.
   - Highlight the overall contribution of the project to business decision-making.
   - Reflect on the lessons learned and potential improvements for future projects.

9. References
    - Include proper in-text citations.
    - Use APA 7th edition format.
    - Minimum FIVE (5) credible references.
    - Two academic papers (journal/conference).

B. Python Implementation Files

This assignment requires two separate submissions:

1. Report Submission (Google Docs)
    - Submit the written report only through Google Classroom, together with a plagiarism detection report.
    - The report must follow the required format (see Section A) and include all visuals, tables, and model outputs to ensure it is self-contained.
    - No need to include code snippets; however, all relevant visuals generated from Python must be shown in the report.

2. Code Submission (ZIP File)
    - Submit a compressed ZIP folder containing all technical files.
    - Naming format: GroupX_RSWY1S2_DataScienceProject.zip
    - The ZIP file must include:
        - All Python files or Jupyter Notebooks (.py / .ipynb)
        - Any prototype files (e.g., Streamlit scripts)

    Compulsory: Include a simple deployment prototype (e.g., Streamlit app, command-line demo, or notebook-based simulation).

3. All visualisations, tables, and model outputs must be included in the report so that it is self-contained.

C. Presentation

Duration: 30 minutes including Q&A.
Format: Group presentation (all members must participate).

Late Submission:
No late assignments will be accepted (get zero). Please **DO NOT** argue with your tutor if you really failed to submit your assignment on time, as the consequence of late submission have been given in advance. However, in certain circumstances, the students may be allowed to turn in the assignment late. With the exception of Extenuating Mitigation Circumstance (EMC) reasons, a penalty for the late submission shall be imposed after submission deadline/extended submission deadline.
The penalty is as follows:

Late submission within 1 - 3 days: Total marks to be deducted are 10 marks.
Late submission within 4 - 7 days: Total marks to be deducted are 20 marks.
Late submission after 7 days: Reject coursework, and zero marks shall be awarded.

Note: For EMC cases, students need to apply for the Leave of Application from the FOCS office as a record of proof of MCs, a Death Certificate. SPM Examination, MUET Examination, etc. are given.

No-Cheating Policy:
A reminder on the no-cheating policy: You are **NOT ALLOWED** to share your work with your peers, but please feel free to discuss with your peers. If cheating is discovered, both parties will take equal blame (get zero). Please note that the assignment should be your own work, although you may incorporate ideas or techniques from books, online resources, etc. Copying materials directly from any sources of materials will lead to zero. You have been warned. Whenever you face any problems, please seek advice from your tutor.