# SCHOOL OF MATHEMATICAL SCIENCES

## ASSIGNMENT FOR B SC (HONS) IN INDUSTRIAL STATISTICS

| | | |
|---|---|---|
| **ACADEMIC SESSION** | **:** | **APRIL 2024 SEMESTER** |
| **SUBJECT** | **:** | **MST3074 MACHINE LEARNING TECHNIQUES FOR DATA MINING** |
| **EXAMINATION** | **:** | **26$^{TH}$ JUNE 2024 - 29$^{TH}$ JULY 2024** |

## INSTRUCTIONS TO CANDIDATES

1. This paper consists of ONE question.
2. Answer ALL questions.
3. This is an **Individual Assignment**, and it will contribute **50%** to your final grade.
4. Complete your modelling using Python.
5. Submit your **final report in PDF and Scripts (.ipynb or .py file)** via eLearn.
6. Submissions with high similarity will receive zero marks.
7. Assignment report must be submitted on their due dates. If an assignment is submitted after its due date, the following penalty will be imposed:

| | |
|---|---|
| • One to two days late | : 20% deducted from the total marks awarded. |
| • Three to five days late | : 40% deducted from the total marks awarded. |
| • More than five days late | : Assignment will not be marked |

# Assignment - Potential Students Prediction

## Introduction

You are working in the marketing team of “Super Star University”. The team is planning on expanding the student base by introducing new bootcamp programs. Currently, the university offers five types of bootcamps: Python, R Programming, Microsoft Excel, Digital Marketing, and Social Media Psychology. Last year's data showed that 18% of prospective students enrolled in these programs. However, identifying potential students was challenging because contacts were made randomly without utilizing available information.

The university now plans to launch a new bootcamp program: “The Super Star Data Science Program”. This time, the university intends to leverage existing and potential student data to target the right candidates.

As a marketing manager at "Super Star University", you are tasked with analyzing student data to provide recommendations to the President and developing a model to predict which prospective students are likely to enroll in this bootcamp. The goal is to make accurate predictions before contacting students.

## Objective:

Develop a model to predict which prospective students are likely to enroll in the newly introduced bootcamp.

## Model:

Students should construct at least two models in order to decide which one is better.

## Data Description

1. **Dataset:** "Student_Data.xls" download from eLearn website.
2. **Variable Descriptions:** Located in the "Description" tab of the Excel file.

## Project Files

1. **Report:** "student_id.pdf" (**Complete analysis** breakdown)
2. **Script**: “student_id.py” or “student_id.ipynb”

Example:

If your ID is 12345, save your report as “12345.pdf” and Python script as “12345.py” or “12345.ipynb” and submit via eLearn.

Ensure your analysis includes:

- Data cleaning and preprocessing steps.
- Exploratory data analysis (EDA).
- Model selection and justification.
- Model training and validation.
- Evaluation metrics for model performance.
- Recommendations based on your findings.

Your report should contains introduction, detail explanations of data preprocessing, your modelling, any analysis involved and clear conclusions.

Table below is the marks distribution:

| Item | Percentage |
|---|---|
| Data Preprocessing | 10% |
| Report: | |
| - Modelling and analysis | 15% |
| - Organization and structure | 15% |
| Python Script | 10% |
| **Total** | **50%** |

**~ END OF PAPER ~**

## Rubric for Case Study

| Criterion | Poor | Moderate | Good | Excellent |
|---|---|---|---|---|
| (10 marks)<br>Data Preprocessing | (0 - 2 marks)<br>The data preprocessing step is missing or very poor. | (3 -5 marks)<br>The data preprocessing steps are not listed and described clearly. | (6 - 8 marks)<br>The data preprocessing steps are relevant, offering details about the data science components. | (9 - 10 marks)<br>The data preprocessing steps introduction offer sufficiently specific details about the data. |
| (15 marks)<br>Report - Modelling and Analysis | (0 - 3 marks)<br>Analytical methods and results are not properly explained.<br>The explanations are not aligned with the data science and business intelligence. | (3 - 7 marks)<br>Analytical methods and results are explained.<br>However, the explanations are confusing, incomplete or lacked relevance to the data science and business intelligence. | (8 - 11 marks)<br>Analytical methods and results are explained.<br>The explanations are appropriate and related to the data science and business intelligence. | (12 - 15 marks)<br>Analytical methods and results are explained well.<br>The explanations are clear, structured, appropriate and related to the data science and business intelligence. |
| (15 marks)<br>Report - Organization and Structure | (0 - 3 marks)<br>The structure of the report is incomprehensible, irrelevant, or confusing. | (3 - 7 marks)<br>The structure of the report is weak.<br>Transition from one to another is weak and sometimes difficult to understand. | (8 - 11 marks)<br>The structure of the report is good.<br>Transition from one to another is smooth. | (12 - 15 marks)<br>The structure of the report is excellent.<br>Transition from one to another is smooth and organized. |
| (10 marks)<br>Python Script | (0 - 2 marks)<br>The script has major error and not able to execute. | (3 - 5 marks)<br>The script has some minor error and able to execute. | (6 - 8 marks)<br>The script has no error and not well documented. | (9 - 10 marks)<br>The script has no error and well documented. |