NGEE ANN POLYTECHNIC

School of InfoComm Technology

Machine Learning

Diploma in Data Science (DS)
Diploma in Information Technology (IT)
October 2023 Semester

INDIVIDUAL ASSIGNMENT 2

(40% of Machine Learning Module)

Deadline for Submission:

Presentation: 28th Jan 2024 (Sunday), 2359 Hours
Report: 8th Feb 2024 (Thursday), 2359 Hours

Student Name :	
Student Number :	

Penalty for late submission:

10% of the marks will be deducted every day after the deadline.
NO submission will be accepted after 15th Feb 2024, 23:59.

MACHINE LEARNING ASSIGNMENT 2

1. OBJECTIVES

In this assignment we will utilize the Machine Learning Models to solve prediction problems using Python:

- To build Machine Learning Models to solve HR Analytics problem and Airbnb problem.
- To train the models by adjusting different hyperparameters.
- To evaluate the model performance and document findings

2. DATASETS

Please refer to assignment 1 document for the problem statements and data dictionaries for HR Analytics problem and Airbnb problem. You have already explored and cleansed the datasets in the previous assignment, so here in this assignment 2, we will simply use the cleansed data for Machine Learning Modeling.

2.1. HR ANALYTICS: hr_data_new.csv

2.2. AIRBNB: listings_new.csv

3. SUGGESTED TASKS

3.1. HR ANALYTICS

You are suggested to tackle this problem in the below FOUR steps.

Step 1: **Load and Sample data**

Load your cleansed dataset (hr_data_new.csv). Perform any additional necessary steps. Provide a short summary on the data cleaning and transformation processes performed.

Step 2: **Build the Machine Learning Model(s)**

Build Machine Learning model(s) using training data to predict whether the employees will be promoted or not. You are encouraged to develop at least 2 different models to solve this problem and contrast between them later.

Step 3: **Evaluate and Improve the Model(s) Performance**

Evaluate the model(s) performance and improve the models' performance by:

- Tuning the model hyperparameters
- Selecting different input features
- Adjusting the input data
- Other effective techniques

Step 4: **Summarize the findings**

Summarize your findings. Recommend the best model and explain why this model performs better than the other models.

3.2. AIRBNB

You are suggested to tackle this problem in the below FOUR steps.

Step 1: **Load and Sample the data**

Load your cleansed dataset (hr_data_new.csv). Perform any additional necessary steps. Provide a short summary on the data cleaning and transformation processes performed.

Step 2: **Build the Regression Model(s)**

Build Machine Learning Model(s) using training data to estimate the listing price. You are encouraged to develop at least 2 different models to solve this problem and contrast between them later.

Step 3: **Evaluate and Improve the Model(s) Performance**

Evaluate the model(s) performance and improve the models' performance by:

- Tuning the model hyperparameters
- Selecting different input features
- Adjusting the input data
- Other effective techniques

Step 4: **Summarize the findings**

Summarize your findings. If you build several different models, please recommend the best model and explain why this model performs better than the other models.

3.3. BONUS [5 MARKS]

Students will deploy their results and models obtained in Section 3.2 on their ***local host computer*** using StreamLit[1] library. Students will be required to give a 10-minute online presentation (through Bongo) on:

- Using Streamlit for visualization of Airbnb dataset.
- Deployment of trained models using Streamlit and
- Your web-based interface.

[1] Refer to the website https://streamlit.io/ for more details.

SUGGESTED REPORT FORMAT & CONTENT GUIDELINES

Write an **INDIVIDUAL** report with the following sections (see Table below). Sample content description is provided for each section. You are free to include other relevant information you deem necessary in the sections. You are strongly encouraged to include screen shots in your explanation, description and analysis.

*(Note: For a page with 1 inch margins, 11 point Calibri font, and minimal spacing elements, a good rule of thumb is **500 words** for a single spaced page)*

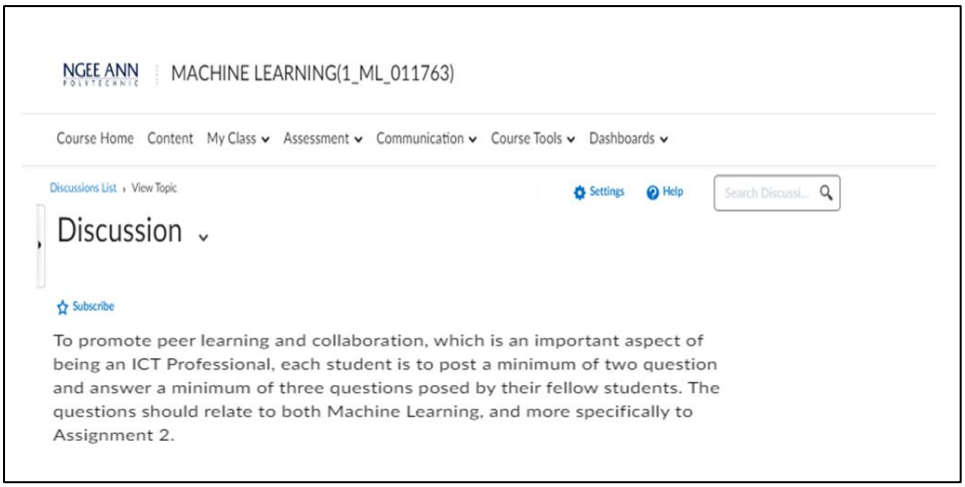
	Suggested Report Sections & Content Guidelines	**Word Count**
1.	Table of Contents	NA
2.	Introduction • A brief introduction on solving classification and regression problems using Machine Learning Models	Min: 200 words Max: 500 words
3.	HR Analytics • Problem understanding and the approaches • Summary of data cleaning and transformation process • Build the model(s) • Evaluate and Improve the model(s) • Summary	Min: 1000 words Max: 4000 words
4.	Airbnb • Problem understanding and the approaches • Summary of data cleaning and transformation process • Build the model(s) • Evaluate and Improve the model(s) • Summary	Min: 1000 words Max: 4000 words
5.	Conclusion • Summarize your work on these two problems	Min: 200 words Max: 500 words
6.	Reflection • Suggest possible further improvement(s) to the current ML solution. • With reference to the module learning objectives stated, reflect on the skills learnt and the skills you could have learnt better.	Min: 500 words Max: 1000 words

4. DELIVERABLES

Discussion

- To promote peer learning and collaboration, which is an important aspect of being an ICT Professional, each student is to post a minimum of two question and answer a minimum of three questions posed by their fellow students. The questions should relate to Machine Learning, and more specifically to Assignment 2.

The questions and answers should be posted in POLITEMALL in the “Discussion” page, as shown below:



You are to print screen your questions and your answers in your report.

Quantity and more importantly, the quality of the questions and answer matters and will be taken into consideration for award of marks.

This portion will account for 10% of the overall grade.

Presentation and demonstration

- Each student will be given **10 minutes** (8 minutes for presentation and 2 minutes for Q&A). Students will present their work in Week 16/ 17 (29Jan to 9Feb). Your tutor will provide detailed information later regarding your presentation slot.

Deliverables

For this assignment, you must submit all the following:

1. A set of **Presentation Slides** in POLITEMall
   - This is the set of presentation slides which you use to conduct your presentation.
   - Deadline for the slides submission is **Sun 28th Jan 2024, 2359 hours**
2. A softcopy **Final Report** via **Turnitin** in POLITEMALL
   - Deadline for report submission is **Thurs 8th Feb 2024, 2359 hours**
3. The **completed "ML_Assignment2.ipynb"** Jupyter Notebook File in POLITEMall
   - Deadline for Jupyter Notebook submission is **Thurs 8th Feb 2024, 2359 hours**

For students attempting the BONUS question, you are to submit:

1. **[BONUS]** *10*-minute online presentation based on the requirements given in Section 3.3. Refer Section 6 for instructions on Video submission.
   - Deadline for Individual VIDEO submission is **Thurs 8th Feb 2024, 2359 hours**

Note: DO NOT PLAGIARIZE (please refer to POLITEMall, Ngee Ann Polytechnic Plagiarism Policy webpage for more information)

5. GRADING CRITERIA

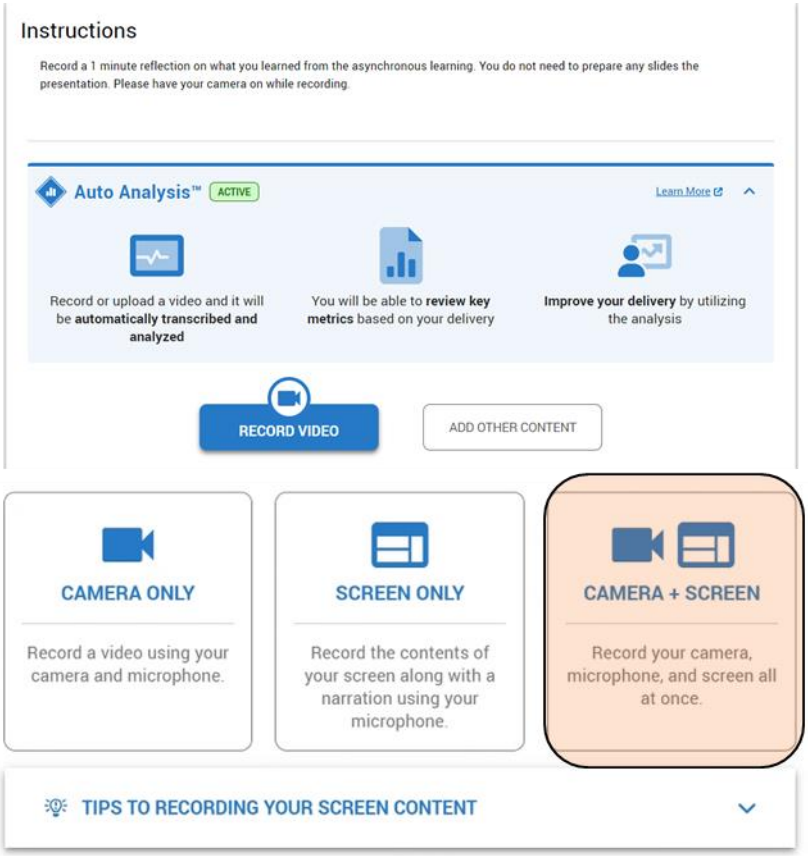
	Grading Criteria	Component Weightage
Discussion	a) Quality of question/answers from forum discussion	**10%**
Presentation	a) Quality of work b) Flow of presentation based on content guidelines (see section 4) c) Quality of presentation slides d) Presentation and articulation skills	**45%**
Final Report	a) Quality of work b) Completeness of report based on suggested report sections and content guidelines (see section 4) c) Clarity of report, Quality of analysis and discussions d) Use of proper visual aids and Use of proper grammar	**45%**

6. VIDEO RECORDED PRESENTATION [*FOR BONUS QUESTION ONLY*]

You are required to do an **online presentation** and share your deployment process and your interface create using StreamLit based on your machine learning model as given in **Section 3.3**. The presentation **should not exceed 10 minutes**. Presentations which exceed the allotted time will be penalized.

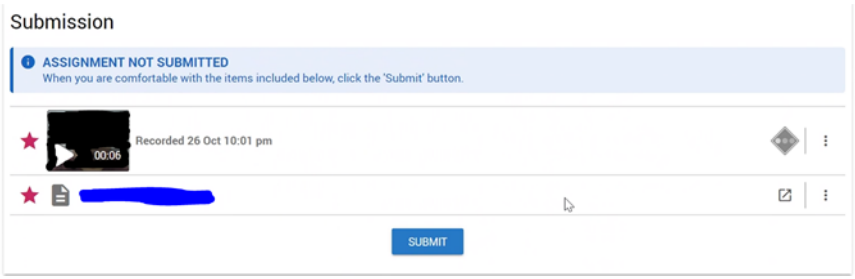
Students will make use of the video assignment app, powered by Bongo, to capture their presentations. Each student is to practice the presentation in advance to ensure completion **within 10 minutes**. The recording must include both webcam (clearly showing the student's face for authentication) and slides or codes (whichever is applicable).

Select the **RECORD VIDEO** option and choose **CAMERA + SCREEN** as shown in the figure below. The figure may differ with the constant update of the Bongo software, hence students may see a different layout but general steps should still apply.



After recording the video, click save (as shown below) and it will be ready for students to append it for submission.

Select the video by clicking on the Star and click **SUBMIT**.



Complete your VIDEO submission no later than **Thursday 8th Feb 2024, 2359 hours** in POLITEMall.