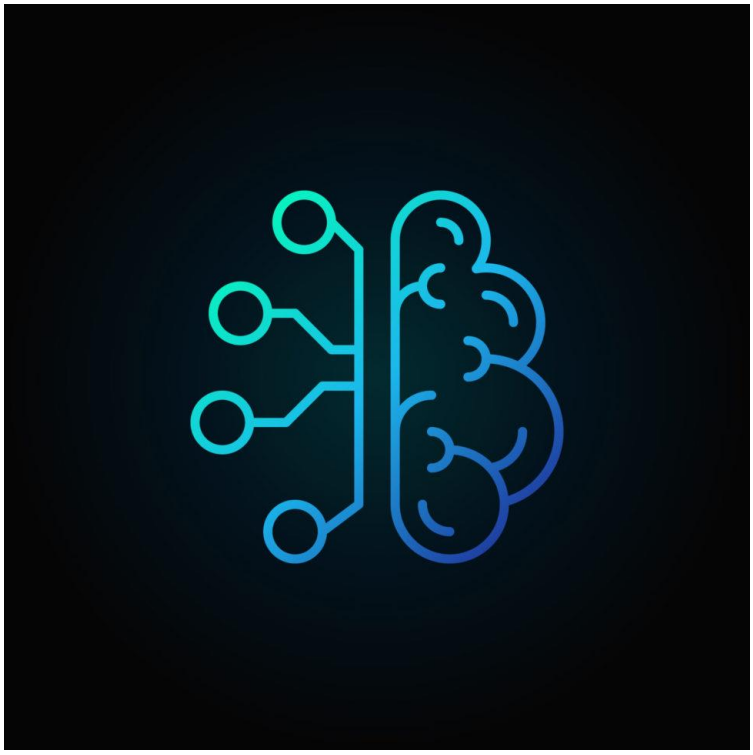

School Name	School of Computing
Semester	AY2223 Semester 1
Course Name	DAAA
Module Code	ST1511
Module Name	AI & Machine Learning

Assignment 2 (CA2: 40%)

The objective of the assignment is to help you gain a better understanding of machine learning tasks of regression and unsupervised learning.

Guidelines

1. You are to work on the problem set individually.

2. In this assignment, you will solve typical machine learning tasks and write a report that describes your solution to the tasks.

3. Write a Jupyter notebook including your code and comments and visualizations. Create a short presentation file (about 10 slides) for your project. Submit your Jupyter notebook, data and the slides in a compressed package (zip file).

4. Students are required to submit their assignment using the assignment link under the Assignment folder. Please remember to include your student name and student admission number on the first page of your assignment report.

5. The normal SP's academic policies on Copyright and Plagiarism applies. Please note that you are to cite all sources. You may refer to the citation guide available at: http://eliser.lib.sp.edu.sg/elsr_website/Html/citation.pdf

Submission Details

Deadline: August 12, 2022, 23:59H
Submit through: BrightSpace

Late Submission

50% of the marks will be deducted for assignments that are received within ONE (1) calendar day after the submission deadline. No marks will be given thereafter.
Exceptions to this policy will be given to students with valid LOA on medical or compassionate grounds. Students in such cases will need to inform the lecturer as soon as reasonably possible. Students are not to assume on their own that their deadline has been extended.

PART A: TIME SERIES (50 marks)

Background

a) Using air-pollution dataset to train a time series model for future air pollution forecasting.
b) You will be given a training dataset to build your time series model, and to make prediction using the test dataset.

Tasks

1. Write the code to solve the time series prediction. For the time series model, use Statsmodels only (do not use other 3rd party libraries such as autoML).
2. Tune the hyperparameters of the time series model to maximize the accuracy for in-sample and out-of-sample prediction.
3. Write a short report detailing your implementation, your experiments and analysis in the Jupyter notebook (along with your python code and comments).
4. Create a set of slides with the highlights of your Jupyter notebook. Explain the time series prediction process, model building and evaluation. Write your conclusions.
5. Using the most optimized model, make a prediction with the testing set, and submit your solution in the Kaggle competition (using the sample_submission.csv template). Remember, before you make the submission on Kaggle, change your Kaggle team name to the format "class-name", such as "2A01-Justin", so that we can identify you.

   https://www.kaggle.com/t/df9b0bf37f6b4c08a1ca6f7422a5ac5e

Submission requirements

1. Submit a zip file containing all the project files (Jupyter notebook), all data sets used, and the slides (PPTX or pdf).
2. Submit online via the Assignment link.
3. Submit your prediction file in the Kaggle competition.

Evaluation criteria:

Background Research & Data Exploration	20%
Modelling and Evaluation	30%
Model Improvement	20%
Demo/Presentation and Quality of report (Jupyter)	20%
Kaggle Competition Evaluation	10%

PART B: UNSUPERVISED LEARNING (40 marks)

Background

You are the HR manager of a big company, and you have some basic data about your employees like Age, Gender, Education, Salary, Performance, Resign status, etc.

Problem Statement

As a HR manager, you want to understand your employees so that appropriate direction can be given to the management to satisfy and retain the employees.

By the end of this case study, you would be able to answer below questions.

- How to achieve employee segmentation using unsupervised machine learning algorithm in Python?
- Describe the characteristics of each employee cluster.
- Which group of employee is the most vulnerable that the management should do something to retain them.

Dataset

Use the Compnay_Employee.csv

Tasks

1. Write the code to solve the clustering task. Use scikit-learn only (do not use other 3rd party libraries).
2. Write a short report detailing your implementation, your experiments and analysis in the Jupyter notebook (along with your python code and comments).
3. Test your clustering with different possible values of k.
4. Determine the best possible value of k. And show how you are able to determine that this is the best value for k.
5. Use more than just one clustering (k-means) algorithm.
6. Create a set slides with the highlights of your Jupyter notebook. Explain the unsupervised machine learning process, model building and evaluation. Write your conclusions.

Submission requirements

1. Submit a zip file containing all the project files (Jupyter notebook), all data sets used, and the slides (PPTX or pdf).
2. Submit online via the Assignment link.

Evaluation criteria:

Background Research & Data Exploration	20%
Feature Engineering	20%
Modelling and Evaluation	20%
Model Improvement	20%
Demo/Presentation and Quality of report (Jupyter)	20%

PART C: Technical Paper (10 marks)

This part of the assignment is to be completed individually. This is a challenge task for students who wish to attempt it for higher marks.

Write a technical paper in single column format on any **ONE** of the following topics.

- Time-Series
- Clustering

The paper should have the following component:

1. Abstract
2. Introduction
3. Related Works
4. Dataset/Methodology/Experiment
5. Discussion
6. Conclusions
7. References

Submit the paper in Word or PDF format (page limit of 10 pages)

— End of Assignment —