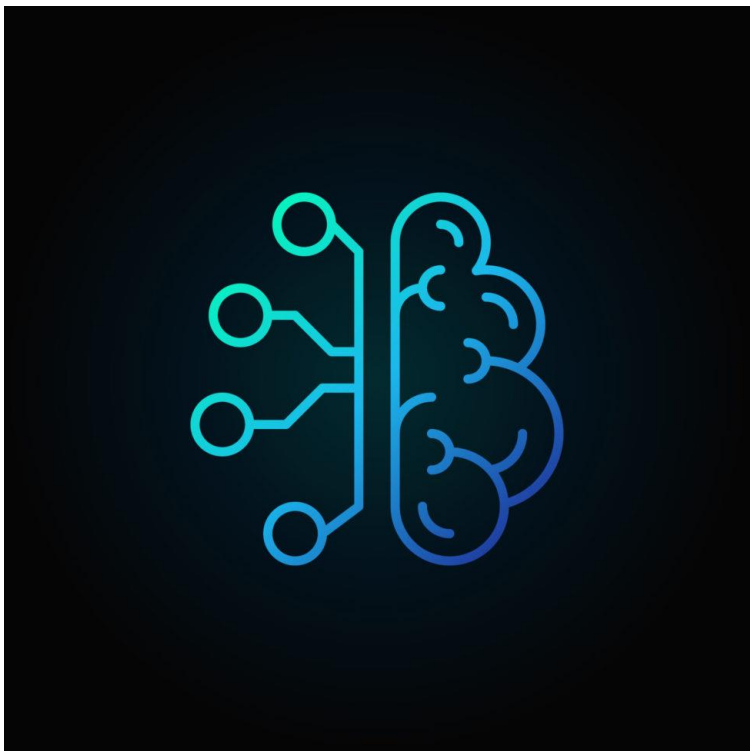

School Name	School of Computing
Semester	AY2324 Semester 2
Course Name	DAAA
Module Code	ST1511
Module Name	AI & Machine Learning

Assignment 1 (CA1: 40%)

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

Guidelines

1. You are to work on the problem sets **individually**.

2. In this assignment, you will solve typical machine learning tasks.

3. Write a Jupyter notebook including your code, data visualization and comments. Create a presentation PowerPoint slides for your work.

4. You are also required to submit the "Declaration of Academic Integrity" form.

5. Submit your Jupyter notebook, PowerPoint slides and Declaration of Academic Integrity form in a zip file. Please name the zip file using this format: **p1234567_WilsonQiu_CA1.zip**.

6. 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 SP academic policy at: https://www.sp.edu.sg/docs/default-source/as-exams/sp_policy-on-use-of-ai-tools-for-academic-work4272ccebf2744e1e9945401cef8f0d2b.pdf?sfvrsn=f00682de_0#:~:text=Academic%20Integrity%20is%20a%20central,copying%20and%20using%20plagiarised%20material.

Submission Details

Deadline: Dec 8, 2023, 23:59pm
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: Predicting Water Quality (50 marks)

This part of the assignment is to be completed individually.

Background

This is a dataset collected from an environmental company. The dataset contains various information on water from different sources. From the perspective of quality control, the company wants to build a machine learning model to predict water quality based on the water properties.

Dataset

You are to use the dataset: **CA1-Classification-Dataset.csv**.

Tasks

1. Write the code to solve the prediction task. You should use scikit-learn to build the machine learning models (no 3rd party libraries).

2. **In the Jupyter notebook**, write your report detailing your implementation, your experiments and analysis (along with your python code and comments). We would like to know:

   - How is your prediction task defined? And what is the meaning of the output variable?
   - Did you process the features in any way?
   - How did you select which learning algorithms to use?
   - Did you try to tune the hyperparameters of the learning algorithm, and in that case how?
   - How do you evaluate the quality of your system?
   - How well does your system compare to a dummy baseline?
   - Is it possible to say something about which features the model considers important? (Whether this is possible depends on the type of classifier you are using)

3. Create a set of slides with the highlights of your Jupyter notebook report. Explain the entire machine learning process you went through, data exploration, data cleaning, feature engineering, model building and evaluation, and model improvement. Write your conclusions. The slides should not exceed 20 pages.

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 B: Predicting Hospital Cost (50 marks)

This part of the assignment is to be completed individually.

Background

This is a dataset to predict the hospital cost in US hospitals based on various patient information, such as ID, Age, Gender, BMI, etc.

Dataset

You are going to use the dataset: **CA1-Regression-Dataset.csv**.

Tasks

1. Write the code to solve the prediction task. You should use scikit-learn for the machine learning models (no 3rd party libraries).

2. **In the Jupyter notebook**, write your report detailing your implementation, your experiments and analysis (along with your python code and comments). We would like to know:

   - How is your prediction task defined? And what is the meaning of the output variable?
   - Did you process the features in any way?
   - How did you select which learning algorithms to use?
   - Did you try to tune the hyperparameters of the learning algorithm, and in that case how?
   - How do you evaluate the quality of your system?
   - How well does your system compare to a dummy baseline?
   - Is it possible to say something about which features the model considers important?

3. Create a set of slides with the highlights of your Jupyter notebook report. Explain the entire machine learning process you went through, data exploration, data cleaning, feature engineering, model building and evaluation, and model improvement. Write your conclusions. The slides should not exceed 20 pages.

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%

— End of Assignment —