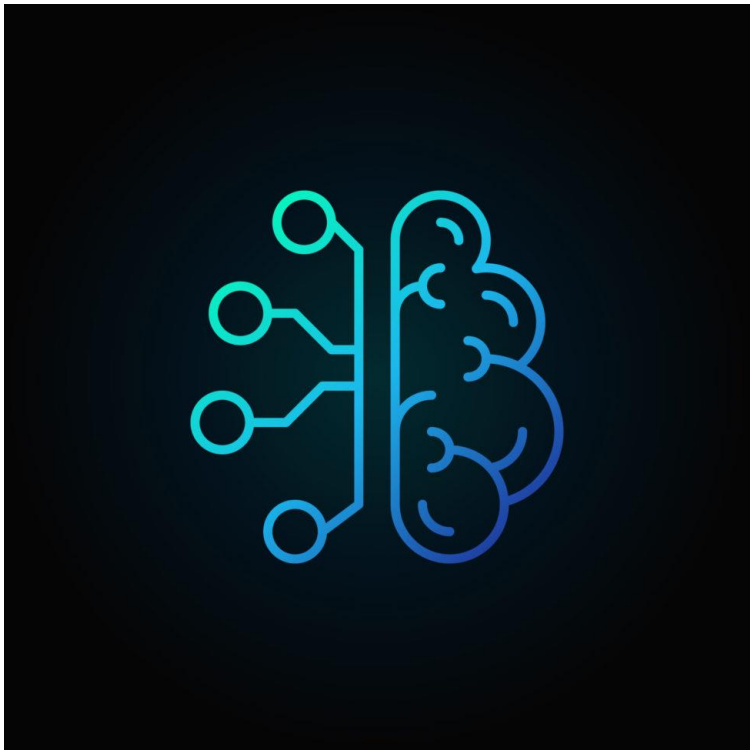

School Name	School of Computing
Semester	AY2021 Semester 1
Course Name	DIT
Module Code	ST0249
Module Name	AI & Machine Learning

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: 2020-08-14 23:59H
Submit through: Blackboard

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: REGRESSION (60 marks)

This part of the assignment is to be completed individually.

Background

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

Dataset

You are to use the dataset.
https://www.kaggle.com/harlfoxem/housesalesprediction

Tasks

1. Write the code to solve the prediction task. Normally you would be using scikit-learn, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
2. Write a report detailing your implementation, your experiments and analysis in the Jupyter notebook (along with your python code and comments). In particular, we'd like to know:

    - How is your prediction task defined? And what is the meaning of the output variable?
    - How do you represent your data as features?
    - Did you process the features in any way?
    - Did you bring in any additional sources of data?
    - 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 stupid baseline?
    - Can you say anything about the errors that the system makes?
    - 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, and model building and evaluation. Write your conclusions.

PART B: UNSUPERVISED LEARNING (40 marks)

Background

a) Given the iris dataset, if we knew that there were k types of iris, but did not have access to a taxonomist to label them: we could try a clustering task: split the observations into well-separated group called clusters.

Dataset

Use the iris dataset from scikit-learn

Tasks

1. Write the code to solve the clustering task. Normally you would be using scikit-learn, but if you'd prefer to work with your own implementation of learning algorithms, or some other toolkit, that is fine.
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 report. 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:

Application of suitable algorithms	20%
Suitable evaluation of algorithms	20%
Background research	20%
Presentation/Demo	20%
Quality of report (Jupyter)	20%

— End of Assignment —