School Name	School of Computing
Semester	AY2021 Semester 1
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.

Guidelines

1. You are to work on the problem sets 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 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 in your notebooks and slides.

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: June 4, 2021 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: CLASSIFICATION (50 marks)

This part of the assignment is to be completed individually.

Background

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). **Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended**. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like ``leaflets three, let it be'' for Poisonous Oak and Ivy.

Dataset

You are to use the dataset.
https://archive.ics.uci.edu/ml/datasets/mushroom

Tasks

1. Write the code to solve the prediction task. You should use scikit-learn ONLY for the machine learning algorithms (do not pip install additional 3rd party machine learning libraries such xgboost, catboost, pycaret, mlbox, auto-sklearn etc).
2. **In the Jupyter notebook**, write your report detailing your implementation, your experiments and analysis (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 hyper parameters 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? For a classification task, you may consider a confusion matrix.
   - 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.

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 B: REGRESSION (40 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. You should use scikit-learn ONLY for the machine learning algorithms (do not pip install additional 3rd party machine learning libraries such xgboost, catboost, pycaret, mlbox, auto-sklearn etc).
2. **In the Jupyter notebook**, write your report detailing your implementation, your experiments and analysis (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, model building and evaluation, and model improvement. 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 on any **ONE** of the following topics.

- Classification
- Regression

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 —