CSCI316 – Big Data Mining Techniques and Implementation
Group Assignment 1
2022 Session 1 (SIM)

10 Marks
Deadline: Refer to the submission link of this assignment on Moodle

<u>One task</u> is included in this assignment. The specification of the task starts in a separate page.

You must implement and run all your Python code in Jupyter Notebook. ***The deliverables include a project presentation, slides and source code.***

All results of your implementation must be reproducible from your submitted Jupyter notebook source files. In addition, the submission must include all execution outputs as well as clear explanation of your implementation algorithms (e.g., in the Markdown format or as comments in your Python codes).

Submission must be done online by using the submission link associated with assignment 1 for this subject on MOODLE. The size limit for all submitted materials is 20MB. DO NOT submit a zip file.

Submissions made after the due time will be assessed as late submissions. Late submissions are counted in full day increments (i.e. 1 minute late counts as a 1 day late submission). There is a 25% penalty for each day after the due date including weekends. The submission site closes four days after the due date. No submission will be accepted after the submission site has closed.

<u>This is a group assignment. Only one submission per group. State the names and student numbers of group members at the beginning of each submitted file.</u>

Marking guidelines

Code: Your Python code will be assessed. The computers in the lab define the standard environment for code development and code execution. Note that the correctness, completeness, efficiency, and results of your executed code will be assessed. Thus, code that produces no useful outputs will receive zero marks. This also means that code that does not run on a computer in the lab would be awarded zero marks or code where none of the core functions produce correct results would be awarded zero marks.

Presentation and explanation: The correctness, completeness and clearness of the project presentation will be assessed.

The Task

(10 marks)

Dataset: Steel Industry Energy Consumption Dataset
(Source: https://archive.ics.uci.edu/ml/datasets/Steel+Industry+Energy+Consumption+Dataset)

The information gathered is from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea. It produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown.

(Reference: Sathishkumar et al.'s Building Research & Information paper in the above link, which is accessible in the UOW digital library, i.e., https://www.uow.edu.au/library/)

Objective
The objective of this task is to develop an end-to-end data mining project by using the Python machine learning library ***Scikit-Learn***. The output of the project is a classification model to predict the *load type*.

Requirements

(1) Main steps of the project are (a) "discover and visualise the data", (b) "prepare the data for machine learning algorithms", (c) "select and train models", (d) "hyperparameter fine-tuning" and (e) "evaluate the outcomes". You can structure the project in your own way. Some steps may be performed more than once.
(2) Clearly explain your findings at each step.
(3) In the steps (c) and (d), select and train at least 3 classifiers (from 3 different algorithms).
(4) Use ~80% data for training and ~20% for testing the models. Stratified sampling *must* be used.
(5) Define some new features by using the User-Defined Transform functionality, which implements a parameter to use those new features or not in the model fine-tuning step (i.e., step (d)).

Deliverables
Deliverables include (1) a presentation of the project in an online tutorial and (2) a submission of the source code and slides via Moodle.