CDS503: Machine Learning
Academic Session: Semester 1, 2022/2023

School of Computer Sciences, USM, Penang

LAB EXERCISE (LAB 2)

Lab 2 Exercise

Look at the Teaching Assistant Evaluation data *(tae.csv)* that is uploaded in eLearn. Look at each attribute and see what type of data it has.

Data Sets

The data set comes from teaching assistant evaluation of the Statistics Department, University of Wilconsin-Madison. The data set is composed of 151 rows of data or examples or instances. Each instance corresponds to a teaching assistant evaluation from a course. Each instance describes features/attributes of an object or entity, which in our case here is a teaching assistant evaluation. In the TEA data set, there are six attributes including the class attribute indicating the class/category information. The six attributes are:

- **Native English speaker or not**
  - 1 (English)
  - 2 (Non-english)
- **Course Instructor**
  - 25 Categories
- **Course (Categorical)**
  - 26 Categories
- **Summer or Regular Semester**
  - 1 (Summer)
  - 2 (Regular)
- **Class Size**
  - Numbers
- **Class Attributes**
  - 1 (Low)
  - 2 (Medium)
  - 3 (High)

Question 1: Do any **pre-processing** to data as *necessary*. Then, answer the following questions:

- What are the **types** of attributes?
  - **Native Speaker**
  - **Course Instructor**
  - **Course**
  - **Semester**
  - **Class size**
  - **Class Attribute**
- Is there any **empty or null** values? What approach you use to address them (remove, replace, etc.)? and why?

Question 2:

Experiment with KNN machine learning algorithm to *predict* what evaluation a teaching assistant (TA) would get based on Teaching Assistant Evaluation data *(tae.csv)*. Use *default* KNN configurations and try **at least** two different values of *k*. Try conduct also with *custom* KNN configurations with **at least** 5-fold cross-validation. Compare the two KNN and specify your findings. Do higher values of *k* lead to better performance? Do cross-validation effect KNN performance?

Post your solution on Lab 02 Submission on **elearn@usm**. Make sure you include your name and lab# on the submission post.

Format: in .ipynb

The due date is **10 November 2022 (5.00pm)**