COMP3055 – 2020 Autumn – Machine Learning Coursework

Deadline: 4pm 24th Dec, 2020
Submit your electronic copy on Moodle

Measures of chlorophyll represent the algal biomass in freshwater that is often used as a proxy for water quality and lake productivity. However, chlorophyll concentrations in lakes are dependent on many variables, such as water fouling, nutrient enrichment, and alterations in climate. We synthesize data of measured Chlorophyll A (CHLA) values with associated factors values, including Temperature and Total Phosphorus (Total P), for the Five Lakes of Maine in the US as below:

Lake Location / Lake Name	Latitude	Longitude
Kezar Lake	44°13'32.9"N	70°53'55.8"W
Cobbosseecontee Lake	44°15'03.1"N	69°56'37.5"W
China Lake	44°25'58.4"N	69°34'10.7"W
Big Lake	45°11'54.4"N	67°38'35.0"W
West Grand Lake	45°14'16.1"N	67°48'05.3"W

Please read excel files for the above "Five Lakes Data" and then make your selection.
https://moodle.nottingham.ac.uk/mod/folder/view.php?id=4764783&forceview=1

Each student should select one lake via "Lake Selection Form":
https://jinshuju.net/f/MzF2HS

After you select one lake, you will perform the following tasks using Python with necessary libraries:

Task-1:
After you select one lake, the lake data from May to October are needed. As for the missing data of the lake, you will use two methods to complete the missing data. One method is "Mean Value", the other method you can choose by yourself. Note that: (1) you cannot copy data from one month before or after; (2) "Median Value" of data from one month before and after is belonging to "Mean Value" method.

Task-2:

After you complete Task-1, you will use five methods to calculate the correlation between CHLA and Temperature & Total P, and then sort the associated factors importance ranking.

Task-3:

Based on your performing Task-1 & Task-2 and findings therein, write a report to compare and analyse different methods: (1) of completing the missing data; (2) of calculating the correlation between CHLA and Temperature & Total P, and then sort the associated factors importance ranking.

Bonus:

As for the above five lakes' data, if you can search and find the matched Total Nitrogen (Total N) and Total Suspended Solid (TSS), you will get a bonus. Note that you should provide the website link where you search and find **the matched** Total N and TSS, and also perform Task-1 & Task-2 & Task-3.

What to submit:

(1) A complete excel file for your selected lake data.
(2) A description on how Task-1 & Task-2 are done.
(3) A report for Task-3. (Word Limit: 800 words)
(4) A zipped file with all your source code. Note that you should properly organize your code with appropriate comments for easy marking and running.

Marking scheme: This coursework takes 30% of your total marks in this module. The marking distribution is given in 100 scaling as follows:

(1) Task-1 (20 marks): Each method to complete the missing data is 10 marks.
(2) Task-2 (30 marks): Each method to calculate the correlation between CHLA and Temperature & Total P, and then sort the associated factors importance ranking is 6 marks.
(3) Task-3 (30 marks)
(4) Coding (20 marks)
(5) Bonus (10 marks): Total N is 5 marks, and TSS is 5 marks.