STAMFORD UNIVERSITY BANGLADESH

Department of Computer Science and Engineering

Midterm Exam, Summer 2022 Semester

MCE 642 : Data Mining

CT: Dr. Md. Tauhid Bin Iqbal

Date and Time: 19/08/2022 and 10:20 AM ~ 12.20 PM

Batch: MCE-S-MIXED **Campus:** Siddeswari

Duration: 2 hours **Full Marks:** 30

---

*(There are **THREE** questions. Answer **ALL**. Numbers in the right indicate corresponding marks.)*

1 (a) Following Tab. 1 represents scholarship info of some MCE/MCA students. [5]

Student	CGPA	IELTS Score	No of Publication	Job Experience	Scholarship
A	3.5	6.5	1	2	Yes
B	2.8	5.5	NULL	1	No
C	3,39	.0	2	3	NULL
D	3.9	12.0	2	2	Yes
E	3.2	6.0	0	-1	No
F	3.6	7.0	3	1	Yes
G	3.6	7.0	2	1	Yes
H	2.9	0.0	0	1	No

Table 1: Scholarship info of the students

(i) Find out the list of entries that needs to be cleaned; also mention the type of their dirtiness (e.g., noisy, missing-data, duplicate-entry etc.).

(ii) Briefly explain how would you clean up those dirty entries. Write down you cleaned entries.

(b) Perform Z-Score normalization for the following samples: [5]

[-34, 21, 13, 46, 999]

2 (a) Two feature columns are given as: [6]

A = [121.1, 90.8, 65.7]

B = [4.9, 8.9, 12.5]

Calculate correlation-coefficient between A and B, and interpret your result from the coefficient value.

(b) Given a data set: [54, 23, 67, 234, 43, 90, 23, 100, 45, 1090, 12, 65] [4]

First, Partition the data using 'equal width partitioning' with 4-intervals, and then smooth them using bin-boundaries.

A = [121.1, 90.8, 65.7]

B = [4.9, 8.9, 12.5]

3 (a) (i) Sample dataset is given as: [5]

[-26, -900, 700, 27, 10000, 16384, 8025, 49, -730, 800, 213, 523]

A data from the above dataset is randomly picked; and, that is: 49. Can you answer which percentile it would fall?

(ii) What is numerosity reduction? What are the differences between parametric and non-parametric numerosity reduction?

(b) Four different lectures are given to same number of students using four(4) different teaching methods. [5]

Students rate each teaching method within a scale of 0~40. Rating scores are illustrated through boxplot in the Fig.1.

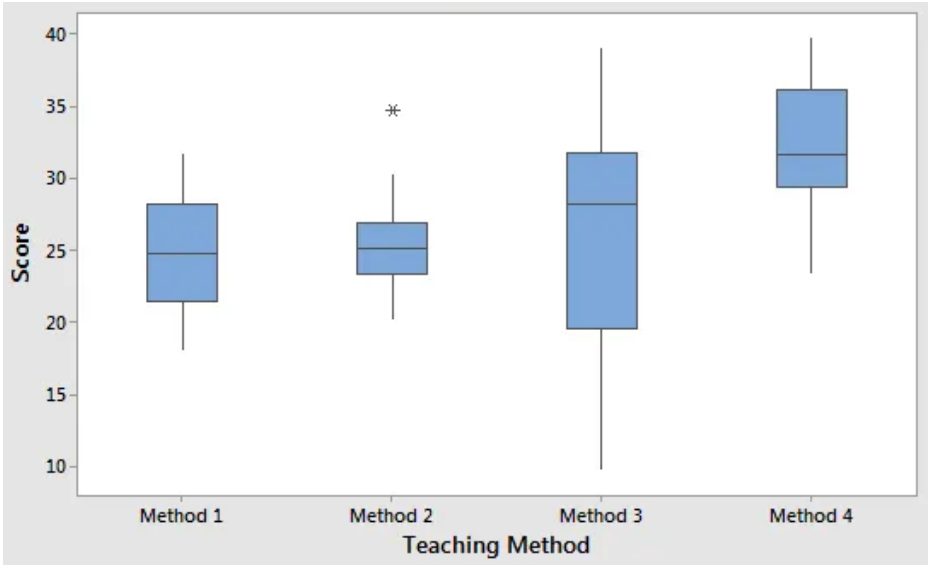

Fig.1: Boxplots for different teaching methods

Share your observations and interpretation from the above boxplots.