VIT
Vellore Institute of Technology

Final Assessment Test – June 2023

Course: MAT5010 - Foundations of Data Science

Class NBR(s): 0506 Slot: A1+TA1

Time: Three Hours Max. Marks: 100

Faculty Name : Prof. SHASHIKIRAN V

KEEPING MOBILE PHONE/SMART WATCH, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE

Answer ALL Questions

(10 X 10 = 100 Marks)

1. a) Define Big Data. What does "volume", "veracity", "variety", and "velocity" for Big Data mean? [6]

   b) What are the types of Data integral to Big Data? [4]

2. Explain briefly different phases of Data Analytics Life Cycle.

3. Calculate the Mean, Median and the mode for the Interval scaled data.

   Marks of the students are grouped and the number of students under each group are given below:

Marks class	10-25	25-40	40-55	55-70	70-85	85-100
Frequency	10	24	48	30	9	4

4. Discuss the significance of first Moment, second moment, third moment and fourth moment in estimating skewness. Derive expression for the same.

5. Find the correlation – coefficient for the following data:

X:	62	64	65	69	70	71	72	74
Y:	126	125	139	145	165	152	181	208

6. Find the partial correlation coefficient $r_{AB.C}$ for the following data.

A	15	18	13	14	19	11	17	20	10	16
B	6	3	8	6	2	3	4	4	5	7
C	25	29	27	24	30	21	26	30	20	25

7. During a country wide investigation, the incidence of T.B was found to be 1%. In a college of 40000 strong 1000 were affected, whereas in another, 120000 strong, 800 were affected. Does this indicate any significance difference?

8. Find the eigenvalues and associated eigenvectors of the matrix

$$\begin{matrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{matrix}$$

9. Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?

10. How does Support vector machine classify given set of data tuples ? SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how to overcome this difficulty and develop a scalable SVM algorithm for efficient SVM classification in large data sets.

⇔⇔⇔