VIT

Vellore Institute of Technology

			SLOT: A1
School of Information Technology and Engineering			
Winter Semester 2022-2023			Continuous Assessment Test – I
Programme Name & Branch	MASTER OF COMPUTER APPLICATION		
Course Code:	MAT 5010	Course Title:	Foundations of Data science
Class Number(s)	VL2022230500506		
Faculty Name(s)	Dr Shashikiran Venkatesha		

Exam Duration: 90 Min. **Maximum Marks: 50**

1. Differentiate Business Intelligence versus Data Science. Illustrate with examples the evolution of analytics from Descriptive to Prescriptive. 10 marks

2. Discuss the Layered approach for Big Data Analysis Framework. 10 marks

3. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) What is the midrange of the data? 2.5 marks

(b) Find the first quartile (Q1) and the third quartile (Q3) of the data? 2.5 marks

(c) Give the five-number summary of the data. 2.5 marks

(d) Draw boxplot for Five number summary data. 2.5 marks

4. Calculate the Mean, Median and the mode for the data given below. 10 marks

The velocities of the jet aircraft were observed at the time of a catapult on an aircraft carrier.

Velocity in knots	135	140	145	150	155	160	165	170
Frequency	4	6	13	8	17	14	7	1

5. Estimate Standard deviation, $\sqrt{\beta_1}$, & β_2 for the following distribution. 10 marks

The Frequency distribution of the heights (in inches) of 200 students in class is given below.

Heights in inches	54	55	56	57	60	61	62	63
Frequency	1	3	7	12	11	34	33	43

VIT
Vellore Institute of Technology

			SLOT: A1
School of Information Technology and Engineering			
Winter Semester 2022-2023			Continuous Assessment Test – II
Programme Name & Branch	MASTER OF COMPUTER APPLICATION		
Course Code:	MAT 5010	Course Title:	Foundations of Data science
Class Number(s)	VL2022230500506		
Faculty Name(s)	Dr Shashikiran Venkatesha		

Open Book Examination.

Exam Duration: 90 Min. **Maximum Marks: 50**

1. A die is thrown 9000 times and a throw of 3 or 4 occurred 3240 times. Is the die unbiased? Find the confidence limits for the probability of getting of 3 or 4.

2. The following data are got from an investigation.

	No of cases	Mean wages	SD of the wage
Sample 1	400	47.4 rupees	3.1 rupees
Sample 2	900	50.3 rupees	3.3 rupees

Find out the two mean wages differ significantly.

3. Obtain the regression lines for the following data. Also determine the co-efficient of correlation.

X	22	26	29	30	31	31	34	35
Y	20	20	21	29	27	24	27	31

4. From the following data, obtain the partial correlation co-efficient $r_{AB.C}$ and $r_{BC.A}$

A	20 15 25 26 28 40 38
B	12 13 16 15 23 15 28
C	13 15 12 16 14 18 14

5. Find the multiple correlation co-efficient $R_{1.23}$ for the following.

1:	50 54 50 56 50 55 52 50 52 51

2: 42 46 45 44 40 45 43 42 41 42,

3: 72 71 73 70 72 72 70 71 75 71

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.

⇔⇔⇔