Dr Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

MIT-WPU

PRN: 1032210755

Mid Semester Examination

Oct 2023

CET2012B - Data Engineering Concepts

Schedule ID: 21573

Faculty/School	Faculty of Engineering and Technology	**Term**	Semester V
Program	TY BTech CSE	**Duration**	**1 Hours 30 Minutes**
Specialization		**Max. Marks**	**50**

Instructions to the Candidate:

1. Write the PRN on the top right-hand corner of the question paper.
2. Draw neat diagrams.
3. Assume suitable data, if necessary.
4. Solve any 5 questions.

Section 1 (5 X 10 Marks)
Answer any 5 questions

1	a) Explain the difference between structured data and unstructured data. Given a set of unstructured data, propose a technique to convert it into structured data. b) Apply the concept of data pre-processing to a real-world dataset by listing the specific steps involved in preparing the data for analysis.	**10 marks** 10	CO1	Applying
2	a) Explain the characteristics and applications of time series data. b) Summarize the role of software engineer, data engineer and a data scientist in IT industry.	**10 marks** 10	CO1, CO2	Remembering
3	a. Describe the Data Engineering Lifecycle Process with the help of a diagram. Illustrate the process with the help of a real-time example. b) Differentiate between discrete data and continuous data. Give any real world example of each type.	**10 marks** 5	CO1, CO2	Applying
4	a) Analyse various reasons of having missing values in a dataset. Explain any one technique to identify and rectify it. b) Elaborate the various data integration techniques in data warehouse.	**10 marks** 5	CO2	Analysing

5	a. Evaluate the binning technique on following dataset using equi-depth, means and boundaries methods using bin depth of 4. The values for data sets are (in ascending order)13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33. b. For the following set of data 1. Draw a scatter plot 2. Calculate Pearson product moment correlation coefficient 3. Asses the statistical significance of your value and interpret your results Students: A B C D E DEC Marks: 1 3 6 10 12 AIES Marks: 5 13 25 41 49	**10 marks** 6	CO1, CO2	Evaluating
6	a. The Table shows the distribution of interest paid to bank investors in a particular year. Draw a histogram to illustrate the data. Interest($): 25- 30- 40- 60- 80- 110- Frequency: 18 55 140 124 96 0 b. Explain following sampling techniques: 1. Random 2. Systematic 3. Stratified 4. Cluster Suppose there are 200 boys and 100 girls in 'Global Talent' school and you need to draw a sample of 60 students. Which sampling techniques could be chosen, explain how this could be done?	**10 marks** 5	CO2	Evaluating

END OF QUESTION PAPER