Register Number:……..

KUMARAGURU
college of technology
character is life

B.TECH. DEGREE EXAMINATIONS: APRIL / MAY 2023

(Regulation 2018)

Second Semester

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

U18AII2206: Introduction to Data Science

COURSE OUTCOMES

CO1: Understand the various aspects of data science and the skill sets necessary for a data scientist.
CO2: Explain the concepts of data storage and Big Data.
CO3: Illustrate the different types of process and tools used in data science.
CO4: Apply the principles of Data Science for analysis using Google Sheets and Excel.

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-
PART A (10 x 2 = 20 Marks)
(Answer not more than 40 words)

1.	Differentiate between data science and big data.	CO1	[K_1]
2.	Why data science is said to be multidisciplinary?	CO1	[K_2]
3.	Write down the significant technical components required for big data analytics to work?	CO2	[K_2]
4.	Find Q2 for the given data: 3, 5, 7, 8, 12, 13, 14, 18, 21	CO3	[K_3]
5.	What is the role of correlation analysis? What are the outcomes of correlation analysis?	CO3	[K_2]
6.	Consider the following runs are scored by Sachin, Dravid and Rohit in a three match ODI series against Pakistan. Compute mean and standard deviation and choose the best player among the three players.	CO3	[K_3]

Sachin	Dravid	Rohit
120	70	210
60	75	0
30	65	0

7.	State the use of Hypothesis testing.	CO4	[K_2]
8.	List out any two data analysis pack.	CO4	[K_1]
9.	What is random sampling?	CO3	[K_2]
10.	Why is normal distribution important?	CO3	[K_2]

Answer any FIVE Questions:-
PART B (5 x 16 = 80 Marks)
(Answer not more than 400 words)

11.	a)	Elaborate the steps involved in the data science pipeline for its life cycle analysis with a neat block diagram.	8	CO1	[K_2]
	b)	Discuss the characteristics of big data and also, explain the different modes and formats of big data.	8	CO1	[K_2]
12.	a)	Explain about computational environments for data scientist.	8	CO1	[K_2]
	b)	Diagrammatically illustrate and discuss the steps involved in the process of Knowledge Discovery from Databases (KDD).	8	CO1	[K2]
13.	a)	On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is 60,849. The actual incomes of the 25 employees are shown below. What are the mean, median, mode, modality, range, midrange, first quartile, third quartile, five-number summary, box-and-whisker plots and outliers of this data? Also, identify if the person is telling you the truth?	8	CO2	[K_3]

12500, 12500, 16430, 17305, 17408,
18980, 20432, 24540, 25676, 28906,
28956, 32654, 33450, 33855, 34983,
35671, 36540, 36853, 37450, 45678,
48980, 94024, 98213, 250921, 478320

	b)	What are the steps involved in Data Transformation for making data suitable for Mining?	8	CO2	[K_2]
14.	a)	The sorted data for price (in rupees) are given as follows: 4, 8, 15, 21, 21, 24, 25, 28, 34. Partition them into three bins by each of the following methods: A) Equal-frequency partitioning B) Smoothing by bin means C) Smoothing by bin boundaries	8	CO2	[K_3]
	b)	Compare Structured and unstructured data. Also describe the challenges with unstructured data.	8	CO2	[K_2]

15. a) A research study was conducted to examine the differences between older and younger adults on perceived life satisfaction. A pilot study was conducted to examine this hypothesis. Ten older adults (over the age of 70) and ten younger adults (between 20 and 30) were give a life satisfaction test (known to have high reliability and validity). Scores on the measure range from 0 to 60 with high scores indicative of high life satisfaction; low scores indicative of low life satisfaction. The data are presented below. Compute the appropriate t-test. 8 CO3 [K_3]

Older Adults	Younger Adults
34	34
22	22
15	15
27	27
37	37
41	41
24	24
19	19
26	26
36	36

b) Discuss about Binomial distribution function and their significance with suitable example scenario. 8 CO3 [K_2]

16. a) Suppose that the data for analysis include 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. Using the data for age, answer the following: 8 CO4 [K_3]

A) Use min-max normalization to transform the value 35 for age onto the range [0.0, 1.0]

B) Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years.

b) The following table shows the heights of sample of Eight fathers and their oldest adult sons. Find correlation coefficient and show that the heights of father and son are positively or negatively correlated 8 CO4 [K_3]

X	165	166	167	167	168	169	170	172
Y	167	168	165	168	172	172	169	171

****