Register Number:……..

KUMARAGURU college of technology
character is life

B.E/B.TECH DEGREE EXAMINATIONS: NOV/DEC 2024

(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.	Define Data Science Life Cycle?	CO1	[K_1]
2.	List down the skills needed to be a data scientist?	CO1	[K_2]
3.	State the difference between Data Augmentation and Data Cleaning?	CO2	[K_2]
4.	Compare and Contrast Structured and Unstructured data?	CO2	[K_2]
5.	Give any four examples for open data and close data?	CO3	[K_2]
6.	What are the process involved in data preprocessing?	CO3	[K_1]
7.	What is data discretization, what are the different types of attributes involved in that?	CO3	[K_2]
8.	What are the different types of distribution functions available?	CO4	[K_2]
9.	Define pmf and cmd?	CO4	[K_1]
10.	State the purpose of Z and T test?	CO4	[K_2]

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

11.	a)	Analyze on how to create a dataset, discuss with a sample case ?	16	CO1	[K_4]
12.	a)	What are the AAAs involved in computational thinking, give an example problem.	8	CO2	[K_2]
	b)	Exemplify the challenges involved in implementing Big data and its counter measures?	8	CO2	[K_3]

13.	a)	How to store and present data in a different format, give example for each case?	16	CO2	[K_2]
14.	a)	The following data (in increasing order) for the attribute age: 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) Use smoothing by bin means to smooth these data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. (b) How might you determine outliers in the data?	8	CO3	[K_3]
	b)	Consider partitioning clustering and the following constraint on clusters: The number of objects in each cluster must be between *n/k(1-delta)* and *n/k (1 + delta)*, where *n* is the total number of objects in the data set, *k* is the number of clusters desired, and delta in (0,1) is a parameter. Can you extend the *k*-means method to handle this constraint? Discuss situations where the constraint is hard and soft.	8	CO3	[K_3]
15.	a)	A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While the line B shows 25 defects out of 600 cars. At α 5%, is the differences between two assembly procedures are significant?	16	CO4	[K_4]
16.	a)	Suppose you start up a company that has developed a drug that is supposed to increase IQ. You know that the standard deviation of IQ in the general population is 15. You test your drug on 36 patients and obtain a mean IQ of 97.65. Using an alpha value of 0.05, is this IQ significantly different than the population mean of 100 (Hint: two tailed z-test)?	12	CO4	[K_4]
	b)	There were 2430 Major League Baseball (MLB) games played in 2009, and the home team won in 54.9% of the games. If we consider the games played in 2009 as a sample of all MLB games, find and interpret a 90% confidence interval for the proportion of games the home team wins in Major League Baseball?	4	CO4	[K_4]

****