Register Number:……..

KUMARAGURU
college of technology
character is life

B.TECH DEGREE EXAMINATIONS: DEC 2022

(Regulation 2018)

Fifth Semester

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

U18AIE5005: Mining Bigdata

COURSE OUTCOMES

CO1: Choose tools to carry out exploratory data analysis and produce effective visualization of given data
CO2: Perform parallel data processing and duplication with Hadoop and Map-Reduce
CO3: Identify suitable data model and algorithms for mining mass data set
CO4: Apply link analysis & mining social network graphs in real time problem

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 interval estimation	CO1	[K_1]
2.	Build a python code to visualize a data set in histogram	CO1	[K_3]
3.	Illustrate row slicing with example	CO1	[K_2]
4.	From the shuffled pack of cards, a card is drawn. This trial is repeated for 400 times, and the suits are given below:	CO2	[K_2]

Suits	Spade	Clubs	Hearts	Diamonds
No of time drawn	120	100	90	90

While a card is tried at random, then what is the probability of getting a spade card

5.	Justify why we need heart message in google file system	CO2	[K_1]
6.	Suppose the selection process of user in big data proportion is 1/10 of the total population. Suggest a suitable solution for sampling the given fixed proportion	CO3	[K_2]
7.	Define Z- testing	CO3	[K_2]
8.	List the various problems in the data stream handling	CO3	[K_1]
9.	Define trust rank	CO4	[K_1]
10.	What is meant by biased random walk?	CO4	[K_1]

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

11.	a)	Explain the 5'v of Big data with suitable example	8	CO1	[K_2]
	b)	Design a python code for the following data visualization i)Heatmap ii)Scatter plot iii)Bar chart iv)count plot	8	CO1	[K_3]
12.	a)	Discuss the various phases of Hadoop file system architecture with suitable diagram	8	CO2	[K_2]
	b)	Consider a file of size 656 MB to be load in the cloud, analyze which file system either GFS or HDFS is suitable to store the file in optimal way. Justify the reason by the following metrics. i) Chunk Partition ii) Reliability iii) Optimal file storage	8	CO2	[K_4]
13.	a)	Suppose you have two matrices with same order m*n. Develop a map reduce framework for the matrix multiplication.	8	CO2	[K_3]
	b)	Perform the following task using Hive Pig Commands i) Create an empty table "Employee(Name, Regno, MailID, MobNo, Address, Designation, salary, GrossPay, Pf, Tax)" to store students details in Hadoop ii) Insert any ten records into the table iii) Arrange the table based on the salary iv) Write a Command to load the table into HDFS	8	CO2	[K_3]
14.	a)	A company wants to improve its sales. The previous sales data indicated that the average sale of 25 salesmen was \$50 per transaction. After training, the recent data showed an average sale of \$80 per transaction. If the standard deviation is \$15, find the t-score. Has the training improved the sales? (α=0.05)	10	CO3	[K_3]
	b)	Many drugs used to treat cancer are expensive. A drug Anthracycline is used in cancer treatment. Treatment costs (in \$) for Anthracycline are provided by a simple random sample of 10 patients.	6	CO3	[K_3]

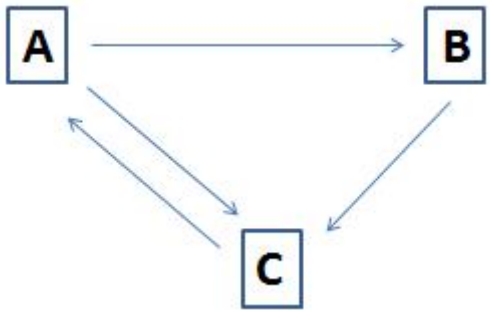
43760 55780 27170 49200 44950

47980 64460 41190 42370 38140

i) Develop a point estimate of the mean cost per treatment with Anthracycline

ii) Develop a point estimate of the standard deviation of the cost per treatment with Anthracycline

15.	a)	Suppose the following data query have been raised from Instagram server, Apply DGIM algorithm to split data stream into buckets stream S= 10010101100010110101010101010110101010101110101010 111010100010110010	6	CO3	[K_3]
	b)	Develop an algorithm/ pseudocode/ program for anomaly detection in network security using Naïve Bayes classifier	10	CO4	[K_2]
16.	a)	Apply AMS algorithm and calculate the second moment of the following data stream, S={a, b, c, b, d, a, c, d, a, b, d, a, b, c, a, a, b}	8	CO4	[K_3]
	b)	Calculate the rank for the following graph using google page rank algorithm (Damping factor = 0.85)	8	CO4	[K_3]



****