Register Number:……………..

KUMARAGURU
college of technology
character is life

B.TECH DEGREE EXAMINATIONS: APRIL / MAY 2023

(Regulation 2018)

Fifth Semester

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

U18AIE5005: Mining Big Data

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. List the 5 v's of Big data CO1 $[K_1]$
2. Develop a python code for 3D Bar Chart CO1 $[K_3]$
3. Compare scatter plot and box plot CO1 $[K_2]$
4. Illustrate namenode in HDFS CO2 $[K_2]$
5. Define shadow master in google file system CO2 $[K_1]$
6. Explain the concept of heart beat message in GFS CO2 $[K_2]$
7. 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 $[K_2]$
8. State moment estimation CO3 $[K_1]$
9. What is meant by biased random walk CO4 $[K_1]$
10. Define trust rank CO4 $[K_1]$

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

11. a) Illustrate the following slicing and indexing operations on any dataset using python or R commands 8 CO1 $[K_2]$
    i) Row Indexing
    ii) Column Indexing
    iii) Slicing using operators

b) Design a python code for the following data visualization 8 CO1 [K$_3$]

i) Heatmap

ii) Scatter plot

iii) Barchat

iv) Count plot

12. Discuss the various phases of google file system architecture with suitable diagram 16 CO2 [K$_2$]

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) Compare HDFS and GFS file system. 8 CO2 [K$_4$]

14. a) Consider the following hypothesis test 10 CO3 [K$_3$]

H0: μ>=1056

Ha: μ<1056

A sample of 400 provided a sample mean of 910. The population standard deviation is 1600.

(i) Compute the value of the test statistic

(ii) What is the p-value?

(iii) At α = 0.05, what is the conclusion?

(iv) What is the rejection rule using the critical value? what is the conclusion?

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

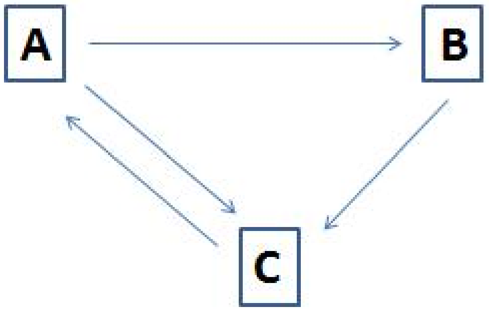
i) Develop as point estimate of the mean cost per treatment with Anthracycline

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

15. Summarize the various issues in big data handling with suitable example. 16 CO3 [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]



****