USN : 1gi20is [illegible] Course Code : 18IS62

Sixth Semester B.E. Semester End Examination, JUNE AUGUST 2023

DATA SCIENCE

Time: 3 hrs. Max. Marks :100

Instructions :1. Answer any FIVE full Questions selecting at least ONE Question from Each Module.

MODULE 1

	L	CO	PO	M
1a. What is Big Data? Explain different Big Data sources.	[2]	[1]	[1]	[6]
1b. Illustrate different data structures used in big data analysis.	[3]	[1]	[1]	[6]
1c. List the different steps of data science life cycle. Briefly explain each one of them.	[2]	[1]	[1]	[8]

OR

2a. Briefly explain the characteristics of big data with examples.	[2]	[1]	[1]	[6]
2b. Discuss characteristics of the Business Intelligence (BI) and compare BI with data science.	[2]	[1]	[1]	[8]
2c. Explain the role of Data Scientist? Illustrate the five main sets of skill sets of Data Scientist.	[3]	[1]	[1]	[6]

MODULE 2

3a. Illustrate the following method of data collection
- i. Observational method of data collection
- ii. Experimental method of data collection

[3] [2] [1] [10]

3b. What is hypothesis testing? If you want to apply hypothesis testing to given data set, choose appropriate steps to be followed. Explain any two hypothesis tests used for analyzing means between two populations. [3] [2] [1] [10]

OR

4a. Explain the following statistical concepts:
- i. Point Estimates
- ii. Confidence interval

[2] [2] [1, 4] [6]

4b. Explain the following with respect to statistics.
- i. Measure of center
- ii. Measure of Variance

[2] [2] [1, 4] [8]

4c. What is exploratory data analysis. Demonstrate how it is helpful in data analytics. [2] [2] [1, 4] [6]

MODULE 3

5a. Explain the concept of clustering and Illustrate with an example. [3] [3] [4] [6]

5b. Analyze the following raw dataset with y as response and x as predictor variables to estimate the two coefficients, β_0 and β_1, using linear regression.

Given: x:1, 2, 3, 4, 5, 6 y:1, 3, 3, 2, 5,6 [4] [3] [4] [8]

5c. What is Linear Regression? Briefly explain any two applications of Linear Regression. [2] [3] [1, 4] [6]

OR

6a. Explain the various steps involved in K means algorithm.

[2] [3] [4] [8]

6b. Discuss the methodology used for identify number of clusters used in k-means clustering algorithm.

[2] [3] [4] [6]

6c. What is logistic regression? Briefly explain the same algorithmically.

[2] [3] [4] [6]

MODULE 4

7a. Discuss main goals of time series analysis. Write a note on any three applications of time series analysis. Write steps followed in Box-Jenkin methodology.

[2] [3] [4] [10]

7b. Illustrate the steps involved in text analysis with example.

[3] [3] [4] [5]

7c. Briefly discuss the ARIMA model of time series analysis.

[2] [3] [4] [5]

OR

8a. Explain the following

i. Auto correlation function
ii. Moving average model

[2] [3] [4] [10]

8b. With Block diagram, discuss the steps involved text analysis.

[2] [3] [4] [10]

MODULE 5

9a. Explain the significance R programming in data science.

[2] [4] [5] [8]

9b. Explain the following concepts of R programming

i. Lists
ii. Vectors
iii. Matrix
iv. Frames
v. Factors
vi. c ().

(Give proper syntax and snippet code)

[3] [4] [5] [12]

OR

10a. Demonstrate the use of rbind and cbind functions in R programming with suitable example.

[3] [4] [5] [7]

10b. Write snippet code to read the following files

i. CSV
ii. Excel

[3] [4] [5] [8]

10c. Explain the summary function of R program.

[2] [4] [5] [5]