Roll No. | | | | | | | | | | |

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. /B.Tech / B. Arch (Full Time) - END SEMESTER EXAMINATIONS, NOV / DEC 2024

INFORMATION TECHNOLOGY
VII Semester
IT5793-DATA SCIENCE AND ANALYSIS
(Regulation2019)

Time:3hrs Max.Marks: 100

CO1	Identify the real world business problems and model with analytics solutions.
CO2	Solve analytics problem with relevant mathematics background knowledge.
CO3	Convert any real world decision making problem to hypothesis and apply suitable statistical testing.
CO4	Write and demonstrate simple applications involving analytics using Hadoop and MapReduce.
CO5	Use open source frameworks for modeling and storing data.
CO6	Perform data analytics and visualization using Python.

BL – Bloom's Taxonomy Levels
(L1-Remembering, L2-Understanding, L3-Applying, L4-Analysing, L5-Evaluating, L6-Creating)

PART- A(10x2=20Marks)
(Answer all Questions)

Q.No.	Questions	Marks	CO	BL
1	What is Big Data Analytics?	2	CO1	L1
2	List the key questions to be answered by all organization steeping into analytics.	2	CO1	L2
3	For the given univariate dataset s={5,10,15,20,25,30} of marks. Find mean, median, mode and standard deviation.	2	CO2	L2
4	The mean salary of the 8 people who work for a small company is 15000 rupees. When an extra worker is taken on this mean drop to 14000 rupees. How much does the new worker earn?	2	CO2	L2
5	Enumerate the drawbacks of regression analysis.	2	CO3	L1
6	Write the MongoDB syntax to find the number of documents in the student collection.	2	CO4	L1
7	List the different types of NoSQL Databases.	2	CO5	L1
8	Why NoSQL is preferred over SQL.	2	CO5	L2
9	How many rows by default the head ( ) command retrieves?	2	CO6	L1
10	Write the correct syntax to access the row in the reverse order.	2	CO6	L1

PART- B(5x 13=65Marks)
(Restrict to a maximum of 2 subdivisions)

Q.No.	Questions	Marks	CO	BL
11 (a) i	Define big data and discuss the challenges in handling big data.	5	CO1	L4
ii	Discuss the various types of digital data, their sources, the challenges associated with handling them, and the advantages and disadvantages of utilizing them.	8	CO1	L3
	OR			
11 (b) i	Define CAP theorem and discuss its implications on distributed system.	5	CO1	L4

ii	Explain the following technologies used in modern data processing and analytics: in-memory analytics, in-database processing, massively parallel processing and shared nothing architecture.	8	CO1	L3
12 (a)	Using Navie Bayes classifier predict whether the person Buys Computer or not whose instance are: age = <=30, income = Medium, Student = Yes, Credit-Rating = Fair based on the following observation.	13	CO2	L4

S.No	Age	Income	Student	Credit - Rating	Buys Computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31-40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31-40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31-40	Medium	No	Excellent	Yes
13	31-40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

OR

12 (b)	The data given is related to gram plant dry weight Y, soil organic matter X1, and kilograms of supplemental soil nitrogen added per 100 square meters X2. Predict the plant dry weight given soil organic matter of 5 and soil nitrogen 4	13	CO2	L4

Y	X1	X2
78.5	7	2.6
74.3	1	2.9
104.3	11	5.6
87.6	11	3.1
95.9	7	5.2
109.2	11	5.5
102.7	3	7.1

13 (a)	Draw the hyperplane that separates the points into positive and negative class using SVM classifier. Points belongs to positive class are (4,1), (4,-1) and (5,0) and points belonging to negative class are (2,0), (0,2) and (0,-2).	13	CO3	L3

OR

13 (b)	Apply K-Means Clustering Algorithm to cluster the data points:	13	CO3	L3

Data points		
A1	185	72
A2	170	56
A3	168	60
A4	179	68
A5	182	72
B6	188	77

		B7	180	71			
		B8	183	84			
		B9	181	88			
		B10	177	76			
	into 2 clusters C1 and C2. The initial centroids of are C1= (185,72) and C2 = (170,56).						
14 (a)	Write in brief about Hadoop Ecosystem, including its core components, associated tools, features, applications, advantages and challenges.				**13**	CO4	L4
	OR						
14 (b)	Explain how MapReduce is used in Matrix-Vector Multiplication, Relational-Algebra Operation, Union, Intersection and Difference Operation				**13**	CO4	L4
15 (a)	Explain in details about Properties, Indexing and Slicing Operation, Arithmetic Operation of NumPy.				**13**	CO6	L3
	OR						
15 (b)	Explain in details about Data Visualization using Pandas and Matplotlib.				**13**	CO6	L3

PART- C(1x 15=15Marks)

(Q.No.16 is compulsory)

Q.No.	Questions	Marks	CO	BL
16.	Using PCA reduce the following two-dimension data into a single dimension data.	**15**	CO5	L5

Features	Ex1	Ex2	Ex3	Ex4	Ex5	Ex6	Ex7	Ex8	Ex9	Ex10
X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
Y	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

CONTROLLER OF EXAMINATIONS ★ ADDITIONAL AU CHENNAI 600 025