RollNo. | | | | | | | | | | |

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

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

INFORMATION TECHNOLOGY

VI Semester

IT5602 - Data Science and Analytics

(Regulation2019)

Time:3hrs Max.Marks: 100

CO1	Identify the real world business problems and model with analytical solutions.
CO2	Solve analytical 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 and 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	Which features of Big data makes it different from traditional data?	2	1	L2
2	List down the skill set required for a Data Scientist.	2	1	L2
3	What are categorical attributes? Give two examples for categorical attributes.	2	2	L2
4	Compute the median and mean for the marks scored by the students studying in 12th grade 87, 53, 47, 67, 39, 97, 80, 72, 32, 94, 25, 39, 80	2	2	L3
5	Differentiate the reasons that lead to model Over-fitting and under-fitting scenarios.	2	3	L2
6	Name some data cleaning methods and why do we clean the data set.	2	3	L1
7	How do you apply grouping and aggregation operation using Map Reduce?	2	4	L2
8	Why do data analyst prefer handling data in HDFS rather than traditional databases?	2	4	L2
9	Write the python statements for training and testing a linear regression model.	2	5	L3
10	What python library/packages are required for building classification models?	2	5	L1

PART- B(5x 13=65Marks)

(Restrict to a maximum of 2 subdivisions)

Q.No.	Questions	Marks	CO	BL
11 (a)	Discuss in detail various sources of Bigdata and present the data analytical life cycle.	13	1	L2
	OR			
11 (b)	Explain about the various categories of data analytics with neat	13	1	L2

	framework along with an illustration.			
12 (a)	(i) State how conditional probability can be defined using Bayes' theorem and apply the same for solving the given scenario: Assume we use NLP techniques to detect spam e-mails in inbox. Suppose that the word 'offer' occurs in 80% of the spam messages in the e-mail account. Also, let's assume 'offer' occurs in 10% of the desired e-mails. If 30% of the received e-mails are considered as a spam and the probability of e-mail with the word 'offer' is 31%, what is the probability that if we receive a new message which contains the word 'offer', it is spam? (ii) Compute the Quartile for the following temperature data. 35, 24, 36, 29, 18, 23, 19, 40, 30, 20, 12.	**8** **5**	2 2	L3 L3
	OR			
12 (b)	Using the data from the table below, calculate variance (X), Standard deviation(X), covariance and correlation X(in years): 5, 3, 4, 10, 15 Y (in Rs): 450, 400, 410, 550, 580	**13**	2 2	L3
13 (a)	The following is a time-series data collected from an academic institution that highlights multiple variables like average hours spent by students for preparation and marks scored by in each semester: **Semester** / **Avg. no. of hrs spent/week towards preparation** / **Marks (%)** 1 / 10 / 60 2 / 20 / 85 3 / 15 / 65 4 / 5 / 50 6 / 25 / 90 7 / 30 / 95 8 / 5 / 40 By applying linear regression modeling identify the relationship between no. of hours spent/week towards preparation and % marks. Also find the standard error. Is this interpolation or extrapolation?	**13**	3 3	L3, L5
	OR			
13 (b)	(i) Write the decision tree construction algorithm. (ii) State the importance of normalization in Data Science. Also normalize the following marks using min-max normalization. Student / Marks A / 67 B / 34 C / 97 D / 76	**8** **5**	3 3	L2 L2, L3

		E	50			
		F	84			
14 (a)	(i) Write a map and reduce function for performing word count from text files distributed across different nodes in a Hadoop cluster. (ii) Write a note on Sharding in MongoDB.			**6** **7**	4 4	L5 L2
	OR					
14 (b)	Explain about Hadoop architecture and the importance of Hadoop Distributed File System.			**13**	4 4	L2
15 (a)	Explain various data visualization techniques and write the pseudocode for the same using Matplotlib assuming a data set.			**13**	5 5	L3
	OR					
15 (b)	Explain the Data Munging and Data pipelining processes in Python Programming.			**13**	5 5	L2

PART- C(1x 15=15Marks)

(Q.No.16 is compulsory)

Q.No.	Questions	Marks	CO	BL
16.	(i) Calculate the Eigen Values and Eigen Vectors for the following matrix. $A = \begin{bmatrix} 2 & 3 \\ 1 & 4 \end{bmatrix}$	**8**		L3
	(ii) Explain the steps for computing the principal components given a matrix of numerical values.	**7**		L2

ADDITIONAL CONTROLLER OF EXAMINATIONS
AU
CHENNAI
600 025