Roll No.

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E / B. Tech (Full Time) END SEMESTER EXAMINATIONS – APRIL/MAY 2024

INFORMATION TECHNOLOGY
VI Semester
IT5602 - Data Science and Analytics
(Regulation 2019)

Time: 3 Hours   Answer ALL Questions   Max. Marks 100

CO 1	Identify the real world business problems and model with analytical solutions.
CO 2	Solve analytical problem with relevant mathematics background knowledge.
CO 3	Convert any real world decision making problem to hypothesis and apply suitable statistical testing.
CO 4	Write and demonstrate simple applications involving analytics using Hadoop and MapReduce.
CO 5	Use open source frameworks for modeling and storing data.
CO 6	Perform data analytics and visualization using Python.

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

PART- A (10 x 2 = 20 Marks)

(Answer all Questions)

Q. No	Questions	Marks	CO	BL
1	List the job roles of a data scientist.	2	CO1	L1
2	How big data is different from traditional data?	2	CO1	L2
3	Draw the histogram and calculate the mode for the age of students studying in 6th grade: 10, 10, 11, 10, 11, 12, 13, 14, 14, 15, 11, 10, 12, 13, 10, 11, 14, 10, 10 State what mode signifies for the given age variable.	2	CO2	L3
4	What is probability density function? State its importance in data analytics.	2	CO2	L2
5	State and analyze the relationship between error, bias, variance metrics of a supervised learning model.	2	CO3	L3
6	Write the various components of each entry in a Common Log Format.	2	CO3	L1
7	What are the support functions for fault tolerance in the Map Reduce framework?	2	CO4	L1
8	Differentiate between ACID and BASE property.	2	CO4	L2
9	Apply matplotlib library and write a simple Python program to plot a linear graph using the plot function with chart title and axis labels.	2	CO6	L3
10	By using any interactive visualization library, write a simple Python program to create an interactive scatter plot.	2	CO5	L3

PART- B (5 x 13 = 65 Marks)

<table>
<tr><th>Q. No</th><th>Questions</th><th>Marks</th><th>CO</th><th>BL</th></tr>
<tr><td>11 (a)</td><td>(i) Consider a big data company that requires a framework to the coordinated processing of programs by several processors working parallel. The processors, each might have their own operating systems and dedicated memory. Also, the processors might work on different parts of the same program. Analyze the given requirement and identify a suitable big data framework that suits best the above-mentioned requirement. Also, draw and brief the architecture by stating how it suits best the given requirement.</td><td>7</td><td>CO1</td><td>L4</td></tr>
<tr><td></td><td>(ii) Compare and contrast the features of structured, semi-structured and unstructured data.</td><td>6</td><td>CO1</td><td>L3</td></tr>
<tr><td colspan="5">(OR)</td></tr>
<tr><td>11 (b)</td><td>(i) Consider an organization that wants to describe or summarise the existing data using existing business intelligence tools and understand why something happened in the past. Identify and brief the apt type(s) of analytics (descriptive, diagnostic, predictive or prescriptive) highlighting its features and relevant open-source tools for the same.</td><td>7</td><td>CO1</td><td>L4</td></tr>
<tr><td></td><td>(ii) Discuss whether or not each of the following activities is a data mining:
1) Predicting the future stock price of a company using historical records.
2) Monitoring seismic waves for earthquake activities.
3) Predicting the outcomes of tossing a (fair) pair of dice.</td><td>6</td><td>CO1</td><td>L3</td></tr>
<tr><td>12 (a)</td><td>(i) Calculate the probabilities for the class labels C1 and C2 and conditional probabilities for the attributes A, B, and C w.r.t the classes C1 and C2 using the following data and represent the same using a conditional probability table:
<table>
<tr><th>RecordID</th><th>A</th><th>B</th><th>C</th><th>Class</th></tr>
<tr><td>1</td><td>0</td><td>0</td><td>0</td><td>C1</td></tr>
<tr><td>2</td><td>0</td><td>0</td><td>1</td><td>C2</td></tr>
<tr><td>3</td><td>0</td><td>1</td><td>1</td><td>C2</td></tr>
<tr><td>4</td><td>0</td><td>1</td><td>1</td><td>C2</td></tr>
<tr><td>5</td><td>0</td><td>0</td><td>1</td><td>C1</td></tr>
<tr><td>6</td><td>1</td><td>0</td><td>1</td><td>C1</td></tr>
<tr><td>7</td><td>1</td><td>0</td><td>1</td><td>C2</td></tr>
<tr><td>8</td><td>1</td><td>0</td><td>1</td><td>C2</td></tr>
<tr><td>9</td><td>1</td><td>1</td><td>1</td><td>C1</td></tr>
<tr><td>10</td><td>1</td><td>0</td><td>1</td><td>C1</td></tr>
</table>
</td><td>7</td><td>CO2</td><td>L3</td></tr>
<tr><td></td><td>(ii) Analyze the following random variables X and Y, denoting the age and weight, respectively using variance and covariance metrics. Consider a random sample of size n = 10 from these two variables.
X = (69, 74, 68, 70, 72, 67, 66, 70, 76, 68)
Y = (153, 175, 155, 135, 172, 150, 115, 137, 200, 130)</td><td>6</td><td>CO2</td><td>L4</td></tr>
</table>

(OR)

12 (b) (i) The following table highlights the marks scored by 100 randomly chosen students from a school. Apply histogram, skewness and kurtosis metrics to analyze the data for its distribution. 7 CO2 L3

Class Mark, x	Frequency, f
61	5
64	18
67	42
70	27
73	8

(ii) Assume a Healthcare company that is collecting customer data in the following structure to perform wellness/health analysis of its customers. Analyze the attributes given in the below table and identify the apt attribute type that the healthcare company has to deal with. Provide suitable justification. 6 CO2 L4

Patient ID	Age (integer)	Age Category (Youth/ Middle Aged/ Senior)	Height (cms)	Weight (kg)	Obesity Status (Yes/No)	Heart Beat Rate	Blood pressure

ADDITIONAL CONTROLLER OF EXAMINATIONS ★ AU CHENNAI 600 025

13 (a) (i) Apply k-means clustering technique on the following data: A1(2,10), A2(2,5), A3(8,4), B1(5,8), B2(7,5), B3(6,4), C1(1,2), C2(4,9). Assume the initial groups as given in the data and the initial centroids for the three groups are A1, B1 and C1 respectively. Use Euclidian distance metric and show the clusters after the 4th iteration. 13 CO3 L3

(OR)

13 (b) (i) Suppose we have to normalize the following data set, 200, 300, 400, 600, 1000 to a new range [0, 1], apply min-max normalization, z-score normalization, and normalization by decimal scaling. Show the normalized values. 7 CO3 L3

(ii) Use real-time applications/case studies to show how the three types of learning techniques are different from each other. 6 CO3 L3

14 (a) (i) Consider a line-counting application on very large distributed document datasets in HDFS environment. Explain the execution of the line counting application as Map and Reduce tasks with necessary diagrams. 13 CO4 CO5 L2

(OR)

14 (b) (i) Explain the process of sharding in MongoDB with the necessary architecture diagram and brief the process of shard indexing. 13 CO4 CO5 L2

15 (a)	(i) Illustrate the significance of various pairwise plots in data analytics with necessary Python codes/functions.	**13**	CO6	L2
(OR)				
15 (b)	(i) Describe the process of data munging and narrate its importance in the data analysis process life cycle with relevant examples.	**13**	CO6	L2

PART- C (1 x 15 = 15 Marks)

(Q.No.16 is compulsory)

Q. No	Questions	Marks	CO	BL
16.	(i) Consider the following data about food intake by cows and milk yield collected from a cattle farm: Food (kg): 4, 6, 10, 12 Milk yield (ltrs): 3.0, 5.5, 6.5, 9.0 The owner of the cattle farm wants to analyze and evaluate the relationship between cows' food intake and milk yield. Identify an apt technique for the same and state the relationship between the variables in the form of a mathematical equation. Also, help the owner predict the milk yield if the food intake is 8 kg.	8	CO3	L5
	(ii) Design and develop a simple application in Python to receive inputs and simulate the above task for the cattle farm. Use Numpy arrays and necessary python libraries for the same.	7	CO6	L6

Food (kg)	4	6	10	12
Milk yield (ltrs)	3.0	5.5	6.5	9.0

----------------ALL THE BEST-----------------