Qn. Set Code-1

Semester: 6th
Programme: B.Tech
Branch: CSE, IT, CSCE, CSSE

SPRING END SEMESTER EXAMINATION-2023

6th Semester B.Tech

DATA ANALYTICS

IT3006

(For 2021 (L.E), 2020 & Previous Admitted Batches)

Time: 3 Hours Full Marks: 50

Answer any SIX questions.
Question paper consists of four SECTIONS i.e. A, B, C and D.
Section A is compulsory.
Attempt minimum one question each from Sections B, C, D.
The figures in the margin indicate full marks.
Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

SECTION-A

1. Answer the following questions. [1 × 10]

(a) Explain the similarity and difference between JSON and BSON with suitable examples.

(b) What is the difference between univariate, bivariate, and multivariate analysis?

(c) Consider the below dataset that contains the number of hours of studies and the actual score received for 3 students in a data set, and the predicted score was calculated with linear regression. Calculate R^2.

#	Number of hrs	Actual score	Predicted score
1	2	74	72
2	3	80	83
3	4	76	79

(d) A time series model is mathematically represented as $Y_t = f(T_t, S_t, C_t, I_t)$ where Y_t is the time series value at time t. T_t, S_t, C_t, and I_t are the trend, seasonal, cyclic and irregular component value at time t respectively. Write the model equation.

(1) When the amplitude of seasonal and irregular variations does not change as the level of trend rises or falls.

(2) When the amplitude of both the seasonal and irregular variations increase as the level of trend rises.

(e) Suppose a hierarchical clustering to be applied in segmenting the students and following sample has been collected. Create the proximity matrix for the below sample. The marks are out of 20 in the mid semester.

Roll No	Sex	Section	Mark
1	Male	CSE -1	10
2	Female	IT – 1	17
3	Male	CSSE – 1	18
4	Female	CSCE - 1	20

(f) Consider the following dataset, wherein TID represents transaction ID and G, A, M represents individual products. In the dataset, 1 represents a transaction that includes the specific products. For instance, TID 1 includes all products and TID 3 includes only M product. Calculate Confidence({G, A} => {M}).

TID	G	A	M
1	1	1	1
2	1	0	1
3	0	0	1
4	0	1	0
5	1	1	1
6	1	1	0

(g) Consider the decagon, which has 10 sides. Three sides are marked 1, two sides are marked 2, one side is marked 3, two sides are marked 4, and two sides are marked 5. Draw a graph representing occurrence of each mark versus its probability.

(h) Consider the following dataset. Let support count be represented with SC. Calculate (SC({E}) + SC({A, B}) + SC({C, D})) / (SC({A, B, C, E}) + SC({A, B, C, D, E}))

Transaction	Itemset
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, B, C, E

(i) A bloom filter with a size of 1000 slots is used to store the information of 100 data stream items using 4 hash functions. Calculate the false positive probability.

(j) What is the probability that a slot is hashed in a bloom filter upon adding n items where n is also the number of slots and k is the number of hash functions?

SECTION-B

2. (a) Consider the following dataset. Draw the MapReduce process to find the number of customers from each city followed by each state, both in the chronological order. [4]

ID	Name	City	State
1	Sujay Lila	Ambikapur	Chhattisgarh
2	Geetha Choudhary	Bhilai	Chhattisgarh
3	Anandi D'Cruz	Bilaspur	Chhattisgarh
4	Surendra Nagarkar	Cuttack	Odisha
5	Balwinder Nagarkar	Bangalore	Karnataka
6	Nitin Nibhanupudi	Mangalore	Karnataka
7	Dinesh Sharma	Cuttack	Odisha
8	Raj Chaudhri	Bilaspur	Chhattisgarh
9	Govind Kumar	Mysore	Karnataka
10	Jayanta Begam	Ambikapur	Chhattisgarh

(b) A retail company wants to enhance their customer experience by analysing the customer reviews for different products, so that they can inform the corresponding vendors and manufacturers about the product defects and shortcomings. You have been tasked to analyse the complaints filed under each product & the total number of complaints filed based on the geography, [4] type of product, etc. You also have to figure out the complaints which have no timely response. Discuss and then model your views concerning descriptive, diagnostic and predictive analytics.

3. (a) In the population, the average IQ is 100 with a standard deviation of 15. A team of scientists want to test a new medication to see if it has either a positive or negative effect on intelligence or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Using hypothesis testing, find the answer to the question i.e., did the medication affect intelligence? The z value (i.e., critical value) from statistical table is found to be 1.96. The solution must mention the null (H_0) and alternative hypotheses (H_a). [4]

(b) Find the relationships of salary between millennials (between the ages of 18 and 34), gen X (between the ages of 35 and 50) and baby boomers (aged 51 and above) of below sample by plotting multiple boxplots in one graph. [4]

Gender	Age	Salary
Male	20	81600
Female	55	61600
Male	38	64300
Female	25	71900
Male	58	76300
Male	45	68200
Female	30	60900
Female	49	78600
Male	60	81700

SECTION-C

4. (a) A consumer electronics company has adopted an aggressive policy to increase sales of a newly launched product. The company has invested in advertisements as well as employed salesmen for increasing sales rapidly. Below dataset presents the sales, the number of [4]

employed salesmen, and advertisement expenditure for 4 randomly selected months. Develop a regression model to predict the impact of advertisement and the number of salesmen on sales.

Month No	1	2	3	4
Sales	5000	5200	5700	6300
Salesmen	25	35	15	27
Advertisement	180	250	150	240

(b) Explain non-linear regression with a suitable example. Subsequently, establish narrate second degree (quadratic), third degree (cubic) and n degree polynomial mathematical model. In general, what techniques applied to determine the right degree of the model? [4]

5. (a) Consider the following dataset consisting of 6 observations that depicts automobile battery sales. Using Simple Exponential Smoothing, calculate the forecasted value of month 7 by calculating smooth observation (S_t) for each month and mean of the squared errors. The smoothing constant is 0.5 and S_1 value is 20. [4]

Month No	Actual
1	20
2	22
3	21
4	18
5	17
6	23

(b) Consider the following dataset capturing monthly sales of actual vs. predicted of an Indian B2C (business to customer) firm. The sales figures are in lakh and presented in INR. [4]

Month No	1	2	3	4
Actual	112	113	122	120
Predicted	113	115	121	119

As a data consultant, the B2C firm hires you for the following and you need to justify your response.

(1) Determine the hybrid error and a hybrid error is determined by 0.3 * MSE + 0.25 * RMSE.

(2) Determine MAPE.

6. (a) Consider the following transactional data in which minimum support is 2 and minimum confidence is 50%. Find frequent itemsets and generate association rules for them by illustrating it with step-by-step process. [4]

Transactions	**List of items**
T1	I1, I2, I3
T2	I2, I3, I4
T3	I4, I5
T4	I1, I2, I4
T5	I1, I2, I3, I5
T6	I1, I2, I3, I4

(b) Consider the following dataset. [4]

Basket	**Product 1**	**Product 2**	**Product 3**
1	Milk	Cheese	
2	Milk	Apple	Cheese
3	Apple	Banana	
4	Milk	Cheese	
5	Apple	Banana	
6	Milk	Cheese	Banana

Calculate Support, Confidence and Lift for the followings:

(1) Apple, Milk

(2) (Apple, Milk) => Cheese

(3) Milk => Cheese

(4) (Apple, Cheese) => Milk

SECTION-D

7. (a) Consider the following hypothetical dataset concerning student characteristics whether or not each student should be hired. Use Naive Bayes Classifier to determine whether or not someone with poor GPA and lots of effort should be hired. [4]

Name	GPA	Effort	Hirable?
Sarah	Poor	Lots	Yes
Dana	Average	Some	No
Alex	Average	Some	No
Annie	Average	Some	Yes
Emily	Excellent	Lots	Yes
Pete	Excellent	Lots	No
John	Excellent	Lots	No
Kathy	Poor	Some	No

(b) Demonstrate a step-by-step process of Agglomerative hierarchical clustering with the following dataset. In addition, illustrate the merge with Dendogram (keep the threshold as 5). Use Manhattan distance for the construction of matrix. [4]

Roll	Mark
1	80
2	90
3	65
4	75
5	95
6	55

8. (a) Design an optimised algorithm for the updation of an element in a Bloom filter. [4]

(b) Consider a Bloom Filter of size 11, with integers as stream elements and two hash functions as follows: [4]

- H1(x) = take odd positioned bits from right in the binary representation of X. Subsequently, treat it as an integer i, and result is i modulo 11.

- H2(x) = same, but take even positioned bits.

(1) Find the filter after the insertion of elements 25, 15 and 35.

(2) Check whether the element y=18 exists in the bloom filter or not. Is it the case of False Positive or False Negative? Explain.

*****