U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

January 2024 Semester End Main Examinations

Programme: B.E. **Semester: VII**
Branch: Institutional Elective **Duration: 3 hrs.**
Course Code: 21CS7OEDAS **Max Marks: 100**
Course: Data Science

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.
3. Use Statistical Tables for appropriate question

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

		UNIT-1	*CO*	*PO*	Marks
1	a)	Brief out the design of data analytics lifecycle? Mention its different phases.	CO1	PO1	**10**
	b)	Consider a scenario where data engineers team needs to analyze a large amount of supply chain application data where they need to assess the quarterly revenue of supply chain company. The data given for the application is raw which contains missing values, outliers and data is not in standardized form. Identify the phase of data analytics life cycle which is applied to process the data. Analyze the above scenario to what steps the data analytics team should follow. Which are the common tools that are applicable in this phase.	CO1	PO1	**10**
		UNIT-2			
2	a)	The specimen of copper wires drawn from a large lot have the following breaking strength (in kg weight) 578, 572, 570, 568, 572, 578, 570, 572, 596, 544. Test using t-test whether the mean breaking strength of the lot may be taken as 578 kg. Calculate t value and corresponding degree of freedom. Compare the t value calculated against 95% confidence interval.	CO1	PO1	**10**
	b)	Explain the concept of Type-I and Type-II error in hypotheses testing.	CO1	PO1	**5**
	c)	Consider the data given below pick a suitable graphics plot which determines the statistical values. Calculate mean, mode, range, IQR, Median for the given data 100, 106, 121, 111, 109, 111, 103, 117, 114.	CO1	PO1	**5**

			UNIT-3			
	3	a)	Explain the process of min-max normalization. Calculate Min_Max normalized value, decimal scaling, z-scale for the following data. Car Name — Weight Datsun 1200 — 1613 Honda Civic Cvcc — 1800 Volkswagen Rabbit — 1825 Renault 5 gtl — 1825 Volkswagen Super Beetle — 1950 Mazda glc 4 — 1985 Ford Pinto — 2046 Plymouth Horizon — 2200 Toyota Corolla — 2265 AMC Spirit dl — 2670 Ford Maverick — 3158 Plymouth Volare Premier v8 — 3940 Dodge d200 — 4382 Pontiac Safari (sw) — 5140	CO3	PO3	**10**
		b)	Explain the process of converting continuous data to categories and converting text to numbers with an example.	CO1	PO1	**10**
			UNIT-4			
	4	a)	Calculate MSB, MSW and F ratio check whether the F ratio value is significant or not (Use F table) for the below data set. Set up an analysis of variance table for the following per acre production. For three varieties of wheat each grown on 4 plots and state if the variety differences are significant. Per Acre Production data Plot of data / Variety of Wheat A: 1: 6; 2: 7; 3: 3; 4: 8 B: 1: 5; 2: 5; 3: 3; 4: 7 C: 1: 5; 2: 4; 3: 3; 4: 4	CO3	PO3	**10**
		b)	Calculate the number of concordant and discordant pairs for the given values. **Observation name** / **Variable X** / **Variable Y** **A**: 1, 2 **B**: 2, 4 **C**: 3, 1 **D**: 4, 3 **E**: 5, 6 **F**: 6, 5 **G**: 7, 7 **H**: 8, 8 **I**: 9, 10 **J**: 10, 9	CO1	PO1	**5**

		c)	For the table of data with values for the x and y variables, find the correlation coefficients. x y 92 6.3 145 7.8 30 3.0 70 5.5 75 6.5 105 5.5 110 6.5 108 8.0 45 4.0 50 5.0 160 7.5 155 9.0 180 8.6 190 10.0 63 4.2 85 4.9 130 6 132 7	CO1	PO1	**5**
			OR			
	5	a)	Illustrate the working of K-means, given the set of values calculate the centroid during each iteration, also group the points to the cluster they belong **Name** | **Variable 1** | **Variable 2** **A** | 7.9 | 8.6 **B** | 6.8 | 8.2 **C** | 8.7 | 9.6 **D** | 6.1 | 7.3 **E** | 1.5 | 2.0 **F** | 3.7 | 4.3	CO3	PO3	**10**
		b)	Generate the frequent 1 item-set, and frequent 2 item-set, for the give transactions with (Min sup=2). **Transaction id** | **Items Brought** 10 | Beer, Nuts, Diaper 20 | Beer, Coffee, Diaper 30 | Beer, Diaper, Eggs 40 | Nuts, Eggs, Milk 50 | Nuts, Coffee, Diaper, Eggs, Milk	CO3	PO3	**5**
		c)	Explain Apriori algorithm for generating frequent item sets.	CO1	PO1	**5**
			UNIT-5			
	6	a)	Explain and illustrate decision tree with a suitable example.	CO1	PO1	**10**
		b)	Differentiate between linear and logistic regression. Given the values age & churned contacts calculate (Y) and probability of churning assume $\beta 0=3.50$, $\beta 1= -0.16$ for age attribute, and $\beta 2=0.38$for Churned contacts using logistic regression.	CO3	PO3	**10**

			Customer Number	**Age(Years)**	**Churned contacts**	**Y**	**Probability of Churning**
			1	50	1		
			2	30	3		
			3	20	1		
			OR				
	7	a)	Explain linear regression and least square technique? Obtain the line of regression equation for the following data using ordinary least square technique. x={1,2,3,4,5,6,7} and Y={9,8,10,12,11,13,14}. Obtain R^2 value and Interpret the model.				CO1 PO1 **10**
		b)	i. Construct Naïve base classifier for the given values. ii. Classify the instance of given attributes of "Red", "Domestic", "SUV" using the constructed model.				CO1 PO1 **10**

Sample	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

******