U.S.N.										

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

Autonomous Institute Affiliated to VTU

April 2024 Semester End Main Examinations

Programme: B.E. **Semester: III**
Branch: Dept. CSE(DS) / AI &DS **Duration: 3 hrs.**
Course Code: 23DS3PCFDS **Max Marks: 100**
Course: Foundations of 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.

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 - I	CO	PO	Marks
1	a)	Elaborate on the various steps of the Data Science process.	CO1	PO1	**10**
	b)	Distinguish between structured and unstructured data with examples.	CO2	PO2	**05**
	c)	Examine the text data given below and classify the type of data as nominal, ordinal, interval and ratio. "Survey results from a group of students were collected to analyze their preferences for extracurricular activities. Each student was asked to choose their favorite activity from a list of options, rank their top three activities, rate their satisfaction on a scale from 1 to 5, and indicate the number of hours per week they spend on their chosen activity."	CO1	PO1	**05**
		UNIT - II			
2	a)	Interpret the need for unequal random sampling. Give an example.	CO2	PO2	**03**
	b)	Elaborate on the five basic steps of the hypothesis test.	CO1	PO1	**07**
	c)	Consider a random sample of 24 friends on Facebook and how many friends they had on Facebook as below. friends = [109, 1017, 1127, 418, 625, 957, 89, 950, 946, 797, 981, 125, 455, 731, 1640, 485, 1309, 472, 1132, 1773, 906, 531, 742, 621] Write a Python program to 1. Find the mean, median, range, and standard deviation(SD). 2. Plot a graph for the obtained SD, SD+mean, and SD-mean. 3. Find the z-score and plot its graph.	C04	PO3	**10**
		OR			

	3	a)	Infer how information gain and entropy are related. Consider the dataset below: i) Calculate the Entropy for the entire data set. ii) Calculate the Information gained for Age, Income, and Credit_rating	*CO3*	*PO3*	**08**
		b)	A new project assignment order is received by an IT company and the authority wants to assign the projects according to the salary package. The dataset is as below. Identify the type and find the probability that 1. Employees that they have undergone professional training 2. An employee has attended professional training and also has a good salary package 3. An employee has a good salary package given that the employee has not undergone professional training	*CO2*	*PO2*	**06**

Age	Income	Student	Credit_rating	Buys_computer
Less than or equal to 30	High	No	Fair	No
Less than or equal to 30	High	No	Excellent	No
Between 31 to 40	High	No	Fair	Yes
Greater than 40	Medium	No	Fair	Yes
Greater than 40	Low	Yes	Fair	Yes
Greater than 40	Low	Yes	Excellent	No
Between 31 to 40	Low	Yes	Excellent	Yes
Less than or equal to 30	Medium	No	Fair	No
Less than or equal to 30	Low	Yes	Fair	Yes
Greater than 40	Medium	Yes	Fair	Yes
Less than or equal to 30	Medium	Yes	Excellent	Yes
Between 31 to 40	Medium	No	Excellent	Yes
Between 31 to 40	High	Yes	Fair	Yes
Greater than 40	Medium	No	Excellent	No

Dataset for Employees

Results		Training		Total
		Without Professional Training	With Professional Training	
Salary package obtained by employees	Poor salary	05	00	05
	Below-average salary	10	00	10
	Average salary	40	10	50
	Good salary	05	30	35
	Excellent salary	00	05	05
	Total	60	45	105

		c)	Distinguish between disjoint and non-disjoint events. Identify the type of event for the below scenarios. i) Coin toss ii) A used car having heated seats and a manual transmission iii) Being a first-year student and being a sophomore iv) Having blue eyes or brown hair	*CO2*	*PO2*	**06**
			UNIT - III			
	4	a)	Provide a explanation of how correlation measures the strength and direction of relationships between two variables, emphasizing both positive and negative correlations.	*CO1*	*PO1*	**05**
		b)	The average rainfall (in mm) for the last 5 years in the state of Karnataka was recorded along with the umbrellas sold in that particular year. **Rainfall (mm)** — **Umbrellas Sold** 121.2 — 52 152.6 — 72 98.4 — 40 171 — 100 85.6 — 34 1. Find the equation of linear regression. 2. Suppose this year the state records a rainfall of 123.2 mm, what will be the expected number of umbrellas that will be sold? 3. Calculate the coefficient of determination (R^2) and comment on it	*CO3*	*PO3*	**10**
		c)	Summarize Receiver Operating Characteristic(ROC).	*CO1*	*PO1*	**05**
			OR			
	5	a)	Depict the various stages of the multinomial logistic regression model with a neat diagram and explain.	*CO1*	*PO1*	**08**
		b)	Suppose, in a dataset, a feature "season" has three values – rainy, summer, and winter. Show how the dummy variable trap can be avoided.	*CO2*	*PO2*	**08**
		c)	Distinguish between linear regression and logistic regression.	*CO2*	*PO2*	**04**
			UNIT - IV			
	6	a)	Distinguish between single and multiple data imputation and write Python code for the same.	*CO4*	*PO3*	**08**
		b)	Elaborate on the common types of inconsistent data.	*CO2*	*PO2*	**06**
		c)	Using the Python's 'pandas' library, write a program to perform data deduplication.	*CO4*	*PO3*	**06**

			UNIT - V			
	7	a)	Discuss the text analytics subtasks.	*CO1*	*PO1*	**08**
		b)	Elaborate on the stages of natural language processing.	*CO2*	*PO2*	**06**
		c)	Write a Python program to create a bag of words using a counter and also remove the stop words using the 'nltk' python library.	*CO4*	*PO3*	**06**

****