U.S.N.										

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

Autonomous Institute Affiliated to VTU

January / February 2025 Semester End Main Examinations

Programme: B.E. **Semester: III**

Branch: 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)	Outline the various steps in the Data Science process with a hierarchical chart.	CO2	PO 2	**12**
	b)	Classify the below list of data into nominal, ordinal, interval or ratio data. (a)ethnic group (b) age (c) family size (d) academic major (e)sexual preference (f) IQ score (g) net worth (h) third-place finish	CO2	PO2	**8**
		OR			
2	a)	Analyze the below paragraph and classify into qualitative and quantitative data. "ShopSmart recently conducted a survey and found that 85% of its customers value personalized recommendations when shopping online. Many customers, particularly in the 25-34 age group, described the shopping experience as "lacking personalization" and noted that the website feels "overwhelming" with too many choices. Sales data reveals that, on average, customers spend about $45 per order, with frequent shoppers placing orders roughly 1.8 times per month. However, over the last six months, customer retention has dropped by 12%, with more	CO2	PO2	**8**

			than 30% of customers not returning after their first purchase. Qualitative feedback from these customers highlights frustration with irrelevant product recommendations and frequent out-of-stock issues. To address these problems, the data science team plans to develop a recommendation system and optimize inventory management based on real-time demand data."			
		b)	Summarize the steps to apply Bonferroni's Principle and its significance in data science process.	*CO3*	*PO3*	**6**
		c)	Discuss the data science Venn diagram.	*CO2*	*PO2*	**6**
			UNIT - II			
	3	a)	For the following data [30 75 79 80 80 105 126 138 149 179 179 191 223 232 232 236 240 242 245 247 254 274 384 470]. Write a python program to compute the mean, median, mode and standard deviation	*CO3*	*PO3*	**10**
		b)	Define hypothesis test. Enumerate the 5 basic steps involved in conducting a hypothesis test. List the 3 most used types of hypothesis test.	*CO2*	*PO2*	**2+5+3 (10)**
			OR			
	4	a)	There are two events: E1: The outcome of a ball delivered cannot be sixer and a wicket E2: A student can get 100 marks in Python coding and 100 marks in R coding. Identify the type of probability each event belongs to. Justify your answer.	*CO2*	*PO2*	**6**
		b)	A factory produces 95% of its products defect-free and 5% defective. 1% of the defect-free products are mistakenly identified as defective by a testing machine, and 2% of the defective products are mistakenly identified as defect-free. What is the probability that a product is defective, given that it has been identified as defective by the machine?	*CO3*	*PO3*	**6**
		c)	A factory produces a set of products, and the prices of the products are as follows (in dollars): Product Prices: [15, 20, 35, 40, 45, 50, 55, 60, 65, 70] Write a program to calculate the standard deviation of the product prices to understand how much the prices vary from the mean price	*CO2*	*PO2*	**8**

			UNIT - III			
	5	a)	You are working as a data analyst at a healthcare company. Your team is tasked with analyzing patient data to gain insights about diabetes risk factors. 1. **Task 1:** Predict the blood sugar level (in mg/dL) of a patient based on their age, BMI, and physical activity. 2. **Task 2:** Predict whether a patient is likely to have diabetes or not (Yes/No) based on the same factors: age, BMI, and physical activity. i) Which type of regression would you use to solve **Task 1**? Explain your choice. ii) Which type of regression would you use to solve **Task 2**? Explain your choice.	*CO2*	*PO2*	**8**
		b)	Illustrate the three types of Logistic regression with examples.	*CO2*	*PO2*	**6**
		c)	Imagine you're working on forecasting future trends based on historical data. Enumerate on the types of time-series models used for prediction, providing examples for each to illustrate their application.	*CO2*	*PO2*	**6**
			OR			
	6	a)	Enumerate on i) Central Limit theorem ii) Bayes Theorem	*CO1*	*PO1*	**10**
		b)	Elaborate in detail the multinomial logistic regression and their stages with suitable diagrams.	*CO2*	*PO2*	**10**
			UNIT - IV			
	7	a)	Suppose you are rolling two dice and want to estimate the probability of getting a total sum of 6. Describe how you would set up a Monte Carlo simulation to solve this problem. Justify the role of random sampling in MC simulation	*CO3*	*PO3*	**12**
		b)	Distinguish between single and multiple data imputation and write Python code for the same.	*CO2*	*PO2*	**8**
			OR			
	8	a)	Implement a simple version of the Fellegi-Sunter model for record linkage. You need to compare two records based on various fields and calculate a matching probability.	*CO4*	*PO4, 5*	**10**
		b)	Elaborate on Entropy based techniques for missing data.	*CO1*	*PO1*	**10**
			UNIT - V			
	9	a)	Discuss the text analytics subtasks.	*CO1*	*PO1*	**10**

		b)	Write a Python program to create a bag of words using a counter and also remove the stop words using the 'nltk' python library.	*CO2*	*PO2*	**10**
			OR			
	10	a)	Highlight the distinctions between text mining and data mining.	*CO2*	*PO2*	**8**
		b)	Write a Python program to lemmatize a list of words using the 'nltk' python library.	*CO3*	*PO3*	**12**

****