U.S.N.										

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

Autonomous Institute Affiliated to VTU

June 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)	Identify and classify as nominal, ordinal interval or ratio data. Justify your answer. i) A survey asks users to rate a restaurant on a scale from 1-10. ii) Measuring the weight of different packages in a shipping warehouse, where the weight has a true zero. iii) Recording the temperatures in different cities throughout the day, where the temperature scale has equal intervals.	*CO2*	*PO2*	**6**
	b)	Implement a python code to calculate the standard deviation for the given temperature data 31,32,32,31,28,31,29	*CO3*	*PO3*	**4**
	c)	A company works on data science projects, one such project is to analyse the precautionary measures for the government to carry out prior to a cyclone. Grievances are collected from social media and other platforms. Discuss the steps involved in this data science process.	*CO1*	*PO1*	**10**
		OR			
2	a)	Enumerate on Data Science Venn diagram and elaborate the role of math and statistical knowledge.	*CO1*	*PO1*	**8**
	b)	Classify the following as structured and unstructured data and justify your answer: i) The data is scientific observations organized in a CSV file ii) The data is collected from facebook posts.	*CO2*	*PO2*	**6**
	c)	Illustrate transformation of data for i. Joining 2 tables ii. Appending data from tables iii. Creating a view	*CO1*	*PO1*	**6**

			UNIT - II			
	3	a)	Define the term Probability. Analyse the statements below and identify the type of probability. i. To find the probability that the card drawn is a heart ii. To find the probability that a card is 4 and red	*CO2*	*PO2*	**5**
		b)	Consider a sample of 5 friends who read books. The frequencies of reading the book are Friends_frequency = [6,4,2,3,5,1,2] Write the python code for i) Mean ii) Median iii) Standard deviation iv) Z-Score v) happiness among friends => happiness = [0.8,0.6,0.4,0.1,0.2,0.3,0.2]. calculate correlation between happiness and Friends_frequency	*CO3*	*PO3*	**10**
		c)	Use the prebuilt python modules to calculate confidence interval. Consider the sample size as 100 and confidence interval as 95 %.	*CO3*	*PO3*	**5**
			OR			
	4	a)	7000 athletes' assembly for a Common wealth games. As part of the survey, we need to know how many athletes take a break to during the games. i) Apply poisson distribution to simulate 3000 employees who take 60 minutes break. ii) Apply poisson distribution to simulate 4000 employees who take 120 minutes break. iii) Apply sample distribution with 250 different point estimates with break times of size 90 each.	*CO1*	*PO1*	**8**
		b)	Elaborate the Hypothesis tests steps involved to query if the sample coffee beans vary significantly in taste from the entire population of beans. Also address the issues involved.	*CO2*	*PO2*	**7**
		c)	A survey conducted categorizes adult BMIs into 4 classes: Normal,over, obsesity and extreme obsesity with the percentages 31.2%, 33.1%, 29.4% and 6.3% respectively. A total of 500 adults were randomly sampled. Calculate the test statistic for Chi -Square Goodness Fit. Note – significance level = 0.05 | | Under/Normal | Over | Obesity | Extreme Obesity | Total | | Observed | 102 | 178 | 186 | 34 | 500 |	*CO1*	*PO1*	**5**
			UNIT - III			
	5	a)	A random sample consists hours students and the marks scored is recorded as x and y respectively. i. Determine the regression equation for the data. ii. Use the regression equation to predict the marks of students in Quiz is who is 5 and 6. iii. Calculate the coefficient of determination (R2) and comment on it.	*CO1*	*PO1*	**7**

			<table><tr><th>Hours studied</th><th>Test Score</th></tr><tr><td>1</td><td>50</td></tr><tr><td>2</td><td>55</td></tr><tr><td>3</td><td>65</td></tr><tr><td>4</td><td>70</td></tr><tr><td>5</td><td>75</td></tr><tr><td>6</td><td>85</td></tr></table>			
	b)	Depict the various stages of the multinomial logistic regression model with a neat diagram and elaborate each stage.	*CO1*	*PO1*	**8**
	c)	Define correlation. A dataset contains person's height (X) and weight (Y) across 5 years. Enumerate on the correlation between X and Y, emphasizing on the type of correlations.	*CO3*	*PO3*	**5**
		OR			
6	a)	Write the python code for Linear regression modelling	*CO3*	*PO3*	**6**
	b)	Derive the equation for logistic regression and identify the category of regression models for the below data samples i) India won the toss. This also implies England lost the toss ii) At a party "Alcohol", "Soft Drinks", "Water" was being served. iii) Students' performance gradually improved from "Better" to "Best".	*CO2*	*PO2*	**8**
	c)	Construct the dummy variable trap for the following tables below: <table><tr><th>Major (k =4)</th></tr><tr><td>Computer Science</td></tr><tr><td>Engineering</td></tr><tr><td>Business</td></tr><tr><td>Literature</td></tr><tr><td>Business</td></tr><tr><td>Engineering</td></tr></table> <table><tr><th>Competition Results</th></tr><tr><td>Win</td></tr><tr><td>Lose</td></tr><tr><td>Lose</td></tr><tr><td>Win</td></tr><tr><td>Lose</td></tr></table>	*CO3*	*PO3*	**6**
		UNIT - IV			
7	a)	Employees that they have undergone professional training and they have rated the course as shown below. Employees are 'A', 'B' and 'C' as shown	*CO3*	*PO3*	**7**

			'A' : [1,2,None,4,5], 'B' : [3,None, 3,1,5], 'C' : [4,5,3,None,5] But the rating contains missing values for few courses. Interpret with python code the approaches the data imputation technique uses to identify the missing values.			
		b)	A dataset contains duplicate entries as shown Name = John, Alice, Bob, John, Age = 25,45,34,25, city = New York, Los Angeles, Los Angles, Chicago . Write the python code to data with duplicates and de-duplicates.	*CO3*	*PO3*	**7**
		c)	Consider a movie data set and demonstrate the process of Singular Value Decomposition.	*CO1*	*PO1*	**6**
			OR			
	8	a)	What are the issues of single imputation? Implement multiple mutation using python code.	*CO3*	*PO3*	**8**
		b)	Illustrate pair wise matching by considering the records of an employee.	*CO1*	*PO1*	**6**
		c)	List any 6 common types of inconsistent data.	*CO1*	*PO1*	**6**
			UNIT - V			
	9	a)	Harry wished they wouldn't, because he was trying to concentrate on finding his way to classes. Harry started to count the staircases. There were a hundred and forty-two staircases at Hogwarts. Apply various text analytics tasks tweet. Also write python code snippets by parsing the tweet i. Tokenization ii. Remove punctuations in the text, using regex iii. Print the stop words from the text iv. Replace the token 'Harry' to 'Harry Potter'. v. Create a Bag of words	*CO3*	*PO3*	**10**
		b)	Relate the role of NLP in applications such as i) Machine Translation ii) Virtual Assistants	*CO1*	*PO1*	**10**
			OR			
	10	a)	A business forum wants to analyze customer reviews to identify common complaints and gauge overall sentiment. Implement the text analytic steps to support the business forum.	*CO1*	*PO1*	**10**
		b)	Implement using python code i. porter stemmer ii. bag of words	*CO3*	*PO3*	**10**

****