U.S.N.										

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

Autonomous Institute Affiliated to VTU

October 2024 Supplementary Examinations

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

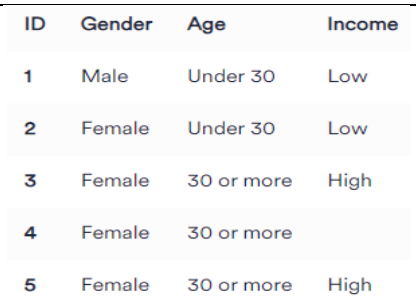
Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

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)	A real estate company estimates the costs for different regions. A dataset has been constructed across 15 years. Depict Venn diagram data science for such an application.	CO1	PO1	06
	b)	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? Identify the type of data that is being analyzed?	CO1	PO1	08
	c)	Define structured and unstructured data. Classify this data into quantitative and qualitative. As of October 31, 2022, Walmart has 10,586 stores and clubs in 24 countries, operating under 46 different names. The company operates under the name Walmart in the United States and Canada. Walmart is the world's largest company by revenue, according to the Fortune Global 500 list in October 2022.	CO2	PO2	06
		UNIT -II			
2	a)	Consider a random sample of 24 friends on Facebook and create a list of how many friends they have on Facebook. Calculate the following 1. Mean 2. Median 3. Z -Score 4. Standard Deviation 5. Plot the visualization of the standard deviation	CO3	PO3	10

		b)	<table><tr><th>day</th><th>outlook</th><th>temp</th><th>humidity</th><th>wind</th><th>play</th></tr><tr><td>D1</td><td>Sunny</td><td>Hot</td><td>High</td><td>Weak</td><td>No</td></tr><tr><td>D2</td><td>Sunny</td><td>Hot</td><td>High</td><td>Strong</td><td>No</td></tr><tr><td>D3</td><td>Overcast</td><td>Hot</td><td>High</td><td>Weak</td><td>Yes</td></tr><tr><td>D4</td><td>Rain</td><td>Mild</td><td>High</td><td>Weak</td><td>Yes</td></tr><tr><td>D5</td><td>Rain</td><td>Cool</td><td>Normal</td><td>Weak</td><td>Yes</td></tr><tr><td>D6</td><td>Rain</td><td>Cool</td><td>Normal</td><td>Strong</td><td>No</td></tr><tr><td>D7</td><td>Overcast</td><td>Cool</td><td>Normal</td><td>Strong</td><td>Yes</td></tr><tr><td>D8</td><td>Sunny</td><td>Mild</td><td>High</td><td>Weak</td><td>No</td></tr><tr><td>D9</td><td>Sunny</td><td>Cool</td><td>Normal</td><td>Weak</td><td>Yes</td></tr><tr><td>D10</td><td>Rain</td><td>Mild</td><td>Normal</td><td>Weak</td><td>Yes</td></tr><tr><td>D11</td><td>Sunny</td><td>Mild</td><td>Normal</td><td>Strong</td><td>Yes</td></tr><tr><td>D12</td><td>Overcast</td><td>Mild</td><td>High</td><td>Strong</td><td>Yes</td></tr><tr><td>D13</td><td>Overcast</td><td>Hot</td><td>Normal</td><td>Weak</td><td>Yes</td></tr><tr><td>D14</td><td>Rain</td><td>Mild</td><td>High</td><td>Strong</td><td>No</td></tr></table> Calculate the information gain for the i. Outlook, temp, humidity and Wind ii. Entropy for the dataset	***CO3***	***PO3***	**10**
			OR			
	3	a)	A software company has 7000 employees. As part of the survey, we need to know how many employees take a break to use the indoor games of the company. 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.	***CO3***	***PO3***	**08**
		b)	Consider the sample size of 1000. Write a python code for i) Sampling Distributions with bins=50. ii) Point Estimate iii) sigma	***CO3***	***PO3***	**06**
		c)	There are records related to the satellite positions which describes observed positions = [102, 178, 186, 34] expected positions = [156, 165.5, 147, 31.5] Write the python code for the Chi-Square test for goodness of fit by finding the degree of freedom.	***CO3***	***PO3***	**06**
			UNIT -III			
	4	a)	Consider the below table for the data related to rainfall and umbrellas sold. i) Create linear regression equation for: The state of Maharashtra experiences rainfall of 123.2 mm what is the expected number of umbrellas sold?	***CO3***	***PO3***	**08**

			ii) Calculate the coefficient of determination (R2) and iii) Write the python code to implement the linear regression. Rainfall (mm) / Umbrellas Sold 121.2 / 52 152.6 / 72 98.4 / 40 171 / 100 85.6 / 34			
	b)	Identify the category of regression models for the below data samples and justify the same for the given data i. The dataset related to patient cancer has target variable malignant/ benign ii. A user ordered these items in a restaurant "Alcohol", "Soft Drinks", "Water". iii. A student has obtained "Good", "Better", and "Best" scores in Computers, Science and Algorithms	***CO2***	***PO2***	**06**
	c)	Demonstrate the stages of the multinomial logistic regression model for a real estate of application that works in Bangalore city. Draw suitable diagrams.	***CO1***	***PO1***	**06**
		OR			
5	a)	A dataset contains with feature_cols and target_variable. feature_cols = ['area_per_square' ,'parking_space', 'facing_site', 'no_rooms'] target_variable = ['Purchased_House'] with the values yes/ no Write python code for i) Linear regression ii) Measure linear regression using MAE, MSE and RMSE.	***CO4***	***PO4***	**10**
	b)	What does ARIMA stand for ? Identify the type of ARIMA time series model for the following i. Financial sector ii. An app to identify the crop to be grown based on the season Also highlight the type of data in above	***CO2***	***PO2***	**10**
		UNIT -IV			
6	a)	What does PCA stand for? A dataset contains F1, F2, … Fn and a target variable. How does PCA identify the correlation and dependencies among the features in a data set? Also write a python code to load the Iris data set and apply PCA using 2 components.	***CO2***	***PO2***	**08**
	b)	Matrix A represents an image.	***CO3***	***PO3***	**07**

			$\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ Reconstruct the original matrix using SVD components. Also write a python code for the same.			
		c)	 ID Gender Age Income 1 Male Under 30 Low 2 Female Under 30 Low 3 Female 30 or more High 4 Female 30 or more 5 Female 30 or more High Given the table contains missing value for income. Using the Missing at random (MAR) predict missing data. Highlight probabilistic relationship approach.	***CO2***	***PO2***	**05**
			UNIT -V			
	7	a)	A sample string "Mary had a little Lamb, little Lamb" write python code for i. Tokenization ii. Replace 'b' with 'p' iii. Print the string after sentence_without_punctuation iv. Istitle() v. lower() Also consider 5 words of your choice to perform stemming.	***CO3***	***PO3***	**08**
		b)	Consider the application which aims at Fraud Detection at an Insurance Company. Identify the text analytics tasks.	***CO2***	***PO2***	**06**
		c)	Google incorporates NLP for the applications such as Autocomplete in Search Engines and voice assistants (Alexa). Discuss the role of NLP in these applications.	***CO2***	***PO2***	**06**

****