Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Final Examination, Fall 2024
Course Code: CSE315, Course Title: Introduction to Data Science
Level:3 Term:2 Batch: 63 & 62

Time: 02:00 Hrs **Marks: 40**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1. a) A quality control officer at a factory tests 10 products randomly from a batch. Each product has a 20% chance of being defective. Find- (i) The probability that exactly 2 products in the sample are defective. (ii) The probability that at most 3 products are defective. — 5

b) Explain the key properties of a normal distribution. Provide examples of real-world that follow a normal distribution. — 2

c) The heights of adult men in a city are normally distributed with a mean of 170 cm and a standard deviation of 8 cm. What is the probability that a randomly selected man is taller than 180 cm? The table is given below for probability selection. — 3

	0.02	0.03	0.04	0.05	0.06
1.1	0.8686	0.8707	0.8728	0.8749	0.8769
1.2	0.8887	0.8906	0.8925	0.8943	0.8961
1.3	0.9065	0.9082	0.9098	0.9114	0.9130

CO2

2. a) A company claims that the average weight of a packaged product is 500 grams. A random sample of 50 packages is taken, and the sample mean is found to be 495 grams with a population standard deviation of 10 grams. Test the company's claim at a significance level of 0.05. **[For a two-tailed test at a significance level of α=0.05, the critical values are: −1.96 (left tail), +1.96 (right tail)]** — 5

b) A fitness trainer wants to determine if a new exercise program leads to weight loss. The weights of 8 participants were recorded before and after the program as follows: — 5

Participant	1	2	3	4	5	6	7	8
Weight before	75	82	68	90	76	85	78	80
Weight after	73	79	65	87	74	82	76	78

At a 5% significance level, test whether the new exercise program significantly reduces weight. **[Critical Value: 1.895]**

CO2

3. a) You are given a dataset of student scores in a class, and tasked with detecting and handling any outliers in the data. The dataset is as follows: — 4

Student ID	1	2	3	4	5	6	7	8
Score	85	90	100	200	88	95	45	92

Write Python code to detect outliers using the **Interquartile Range (IQR)** method. After detecting the outliers, explain and write code how you would handle them.

CO3

	b)	Write down techniques of normalization in feature scaling.	2	
	c)	Feature extraction is a key step in many machine-learning workflows. In the context of supervised learning, explain the following: 1. Explain **feature extraction**, and why it is important in machine learning models. 2. Discuss at least two common techniques used for feature extraction, providing a brief description of each. 3. How does feature extraction improve the performance of a machine learning model?	4	

4.	a)	`data = {`
		`'Student': [1, 2, 3, 4, 5],`
		`'Math': [85, 90, 78, 92, 75],`
		`'Science': [78, 88, 74, 85, 80],`
		`'English': [92, 76, 80, 89, 82],`
		`'Total': [255, 254, 232, 266, 237],`
		`'Average': [85.0, 84.67, 77.33, 88.67, 79.0]`
		`}`
		i. Write Python code to transpose the dataset such that each row represents a subject (or calculated field, e.g., Total or Average), and each column represents the scores of the 5 students.
		ii. Drop the "Total" and "Average" rows from the transposed data.
		iii. Convert the transposed DataFrame into a new format where the column names are "Subject" and the scores of all 5 students are stored as a list. Display the final transformed DataFrame.
			4	

	b)	A company is analyzing how the number of hours spent on training (X_1) and the number of years of experience (X_2) of employees affect their productivity score (Y). The following data is collected from three employees: Hours of Training (X1) | 10 | 15 | 20 Year of Experiences (X2) | 2 | 3 | 4 Productivity Score (Y) | 50 | 60 | 70 i. Write the model equation in matrix form. ii. Use the normal equation $(X^T X)^{-1} X^T Y$ to calculate the values of the regression coefficients β_0, β_1, and β_2. iii. Interpret the meaning of β_1, and β_2 in the context of the data. What will be the Productivity Score if the employee has 6 years of experiences and done 12 hours of training?	6	CO3