



# 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	CO2																									
	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																										
	<table><tr><td></td><td>0.02</td><td>0.03</td><td>0.04</td><td>0.05</td><td>0.06</td></tr><tr><td>1.1</td><td>0.8686</td><td>0.8707</td><td>0.8728</td><td>0.8749</td><td>0.8769</td></tr><tr><td>1.2</td><td>0.8887</td><td>0.8906</td><td>0.8925</td><td>0.8943</td><td>0.8961</td></tr><tr><td>1.3</td><td>0.9065</td><td>0.9082</td><td>0.9098</td><td>0.9114</td><td>0.9130</td></tr></table>						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			
	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																												
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 $\alpha=0.05$ , the critical values are: -1.96 (left tail), +1.96 (right tail)]					5	CO2																									
	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: <table><tr><td>Participant</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr><tr><td>Weight before</td><td>75</td><td>82</td><td>68</td><td>90</td><td>76</td><td>85</td><td>78</td><td>80</td></tr><tr><td>Weight after</td><td>73</td><td>79</td><td>65</td><td>87</td><td>74</td><td>82</td><td>76</td><td>78</td></tr></table> At a 5% significance level, test whether the new exercise program significantly reduces weight. [Critical Value: 1.895]					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
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																									
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: <table><tr><td>Student ID</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr><tr><td>Score</td><td>85</td><td>90</td><td>100</td><td>200</td><td>88</td><td>95</td><td>45</td><td>92</td></tr></table> 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.					Student ID	1	2	3	4	5	6	7	8	Score	85	90	100	200	88	95	45	92	4	CO3							
Student ID	1	2	3	4	5	6	7	8																									
Score	85	90	100	200	88	95	45	92																									

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 <b>feature extraction</b> , 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)	<pre>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] }</pre> 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	CO3											
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:  <table border="1"><tr><td>Hours of Training (<math>X_1</math>)</td><td>10</td><td>15</td><td>20</td></tr><tr><td>Year of Experiences (<math>X_2</math>)</td><td>2</td><td>3</td><td>4</td></tr><tr><td>Productivity Score (<math>Y</math>)</td><td>50</td><td>60</td><td>70</td></tr></table>  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 $\beta_0$ , $\beta_1$ , and $\beta_2$ . iii. Interpret the meaning of $\beta_1$ , and $\beta_2$ 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?	Hours of Training ( $X_1$ )		10	15	20	Year of Experiences ( $X_2$ )	2	3	4	Productivity Score ( $Y$ )	50	60	70
Hours of Training ( $X_1$ )	10	15	20											
Year of Experiences ( $X_2$ )	2	3	4											
Productivity Score ( $Y$ )	50	60	70											