Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Final Examination, Spring 2026
Course Code: 328, Course Title: Introduction to Data Science
Level: 3 Term: 3 Batch: 65

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.]

Read the following scenario carefully and answer to the question no 1 to 3 using this scenario.

On their first day at *CerebralStation65*, five graduates from Daffodil International University, Sushmita, Arman, Jannat, Noman, and Mahim were assigned to a collaborative project in different feature engineering tasks. The company provided the 5 graduates with a cleaned dataset (The first 5 rows of the whole dataset is given below) and divided the tasks among the team members. Sushmita was asked to create new variables from the existing data, making the dataset more informative. Arman worked on adjusting the numerical values so that they were placed on a common scale, ensuring fairness in the model. Jannati focused on reducing the complexity of the dataset by extracting features into a smaller set of meaningful components. Noman converted categorical information into numerical form so that it could be processed by algorithms. Mahim concentrated on identifying the most relevant attributes from the expanded dataset, removing those that added little value.

Employee	Age	Salary	Department	Projects	Device
A	24	30000	IT	2	MOBILE
B	29	45000	HR	3	TABLET
C	35	60000	FINANCE	5	MOBILE
D	40	80000	MARKETIN G	4	LAPTOP
E	31	50000	IT	6	TABLET

1.	*a)*	**Discuss** one preprocessing technique to handle any identified issue (e.g., outlier, scaling, skewness). Justify your choice.	[2]	CO1
	b)	**Describe** appropriate visualization techniques to understand: • The relationship between Age and Salary • The distribution of Salary • The relationship between Department and Projects • Device usage For each selected visualization: 1. Name the visualization technique 2. Justify why it is appropriate	[8]	

		3. Describe one meaningful pattern, trend, or insight that can be observed from the above dataset.		
2.	*a)*	As a friend of Noman, which technique would you suggest him to **apply** to complete his task? **Illustrate** pros and cons of the technique.	[3]	CO2
	b)	**Apply** the technique visually and show how the updated dataset looks like?	[4]	
	c)	Arman was asked to keep the age attribute in the range of 1 to 7. **Apply** appropriate method to convert the data for Employee A and D mentioning the name of the method chosen by you.	[3]	
3.	*a)*	**Interpret** whether predicting Salary from Age and Projects is a classification or regression task, with justification. **Illustrate** the dependent and independent variables for this prediction task.	[3]	CO2
	b)	The company defines a new categorical variable based on Salary: Salary > 50,000 → High Salary (1) Salary ≤ 50,000 → Low Salary (0) **Interpret** the new column named Class and update the dataset accordingly.	[2]	
	c)	A new employee F has the attributes, Age = 30 and Projects = 4. **Apply** K-Nearest Neighbors (K = 3) to predict whether Employee F belongs to the High Salary (1) or Low Salary (0) group with justification. Consider only Age and Projects to find the three nearest employees (show distance comparison).	[5]	
4.		"Bangladesh Bank" is developing a predictive machine learning model to detect fraudulent credit card transactions. Fraud detection is a critical real-world problem where the dataset is highly imbalanced (i.e., thousands of legitimate transactions happen for every single fraudulent one). The data science team tested 10,000 transactions, containing exactly 150 actual fraudulent transactions and 9,850 legitimate transactions. They trained two models and generated the following results: **Model X (KNN):** Confusion Matrix: True Positives (100), False Negatives (50), True Negatives (9000), False Positives (850). Performance Metrics: Accuracy: 91.0%, Precision: 10.5%, Recall: 66.7% **Model Y (Logistic Regression):** Confusion Matrix: True Positives (130), False Negatives (20), True Negatives (9650), False Positives (200). Performance Metrics: Accuracy: 97.8%, Precision: 39.4%, Recall: 86.7%		CO3
	a)	**Analyze** why relying solely on the high Accuracy scores (91.0% and 97.8%) is misleading for this specific real-world problem, and explain why it gives a false sense of security.	[3]	
	b)	**Analyze** the specific meaning of Precision and Recall strictly in the context of detecting credit card fraud. Using these metrics, compare how effectively Model X and Model Y isolate the actual fraudulent transactions.	[3]	
	c)	In the banking industry, a False Negative (missing a fraudulent transaction) leads to direct financial loss, while a False Positive (blocking a legitimate user's card) leads to severe customer dissatisfaction. **Evaluate** both models against these competing business constraints. Deduce which model the bank should deploy and justify your decision.	[4]	