ourdiu.com | Shared by SUPAN ROY

Daffodil International University

Faculty of Science & Information Technology

Department of Computer Science & Engineering

Midterm Examination, Spring 2026

Course Code: 328, Course Title: Introduction to Data Science

Level:3 Term:3 Batch: 65

Time: 01:30 Hrs **Marks: 25**

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)*	**Explain** the correct sequence of data science steps a bank would follow to predict fraudulent transactions using customer transaction histories, profiles, and device login data. Then justify why each step logically follows the previous one. Steps: A) Evaluate the model performance using metrics like accuracy and F1-score B) Collect transaction and customer data C) Communicate insights to the risk management team D) Clean and normalize the data E) Train a machine learning model to detect fraudulent transactions F) Explore the data to find trends and suspicious patterns	[5]	**CO1**
	b)	**Explain** how an Agri-Tech company could design a crop disease prediction system using real-time IoT soil sensor data and daily drone images by describing a suitable data storage architecture that allows the system to continue operating even when data formats change, and by clarifying which Big Data challenges (the 5 V's) are relevant to each data source with justification for your reasoning.		
2.	*a)*	**Apply** your knowledge of variable types and measurement scales to classify the following variables for each patient, and record the measurement scale used for each variable.	[5]	**CO2**
3.		Three colleagues, Turjo, Arunob, and Tanvir, are collaborating on an analytical research project concerning the Rohingya issue. From the outset, they recognized that collecting comprehensive information on the Rohingya population would be challenging. Consequently, they decided to employ sampling techniques to obtain representative data. i. Turjo proposed dividing the population into five distinct age groups: 0–20, 21–40, 41–60, 61–80, and 81 and above. From each group, participants would be selected to ensure age-based representation.	[5]	

Patient ID	Age	Weight	Blood Pressure	Blood Group	Cholesterol	Smoker	Exercise
P201	50	82	145/95	B+	220	Yes	Low
P202	28	58	115/75	O+	155	No	High

ourdiu.com | Shared by SUPAN ROY

		ii. Tanvir suggested identifying the geographic locations in Bangladesh where Rohingyas reside, segmenting these areas into different regions, and then selecting participants from each region. iii. Arunob, however, emphasized that many Rohingyas may conceal their identity, making it difficult to identify them directly. He therefore recommended using a referral-based approach: beginning with Rohingyas known to the researchers, who would then introduce additional participants, thereby expanding the sample through successive referrals.		
	a)	**Apply** appropriate sampling techniques to collect representative data on the Rohingya population for a research study by selecting one or more of the proposed methods (age-group sampling, geographic-region sampling, or referral-based sampling) and describing in practice how you would select participants while accounting for practical constraints such as accessibility and representativeness.		
4.		Prior to the upcoming World Cup, Bangladesh's leading batsman, Saif Hasan, had shown remarkable consistency in One Day Internationals (ODIs). In the twelve matches before the ODI series against Australia, he had never scored fewer than fifty runs. However, in Bangladesh's final series before the World Cup, his performance declined, with scores of 0, 4, and 0. The sequence of his last fifteen ODI scores is: 51, 100, 56, 59, 160, 76, 54, 100, 90, 89, 81, 98, 0, 4, 0. A Bangladeshi cricket selector believes that one or two bad innings should not overshadow his overall ability, and wants to assess his performance statistically before the World Cup.	**[10]**	**CO3**
	a)	Calculate the mean and median for Saif Hasan's 15 innings. **Analyze** the difference between these two measures in the context of his recent performance slump. Based on the distribution of the scores, which statistic provides the most reliable indicator of his "true" form for the World Cup selectors, and why?		
	b)	Construct a boxplot for the 15 innings. Using the Interquartile Range (IQR) method, mathematically **analyze** the dataset to determine if any of the recent low scores (0, 4, 0) qualify as formal statistical outliers. Explain how the overall skewness of the distribution either supports or contradicts the selector's belief that the recent bad innings are just an anomaly.		
	c)	The selector believes Saif's true batting average remains strong despite recent matches, claiming his true mean is above 50. A hypothesis test is conducted (H_0: $\mu = 50$ versus H_a: $\mu > 50$) at a 5% significance level, yielding a p-value of 0.03. **Analyze** this statistical result by applying the decision rule to the null hypothesis, and explain exactly what this outcome implies for the selector's confidence in picking Saif for the World Cup.		