Daffodil International University

Faculty of Science & Information Technology

Department of Software Engineering

Midterm Examination, Spring 2026

Course Code: DS 331; Course Title: Introduction to Data Science and Data Management & Analysis

Sections & Teachers: ALL

Time: 1 Hour 30 Mins **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)***	Consider the following scenario: A mid-sized e-commerce company is planning to build a customer churn prediction system. The company has historical data including customer demographics, purchase history, website activity, and customer support interactions. However, previous data projects in the company failed due to poor planning, weak data preprocessing, and lack of proper evaluation. As a Data Science consultant, **analyze** the complete Data Science pipeline for developing the churn prediction system. Examine how each stage contributes to the project and **analyze** how the stages are interconnected in determining the overall outcome.	**4**	**CLO-1** ***Level-4***
	b)	Consider the following scenario: A healthcare analytics company wants to develop a predictive model to identify high-risk patients using clinical records, lab reports, and lifestyle data. The company plans to use Python for the entire workflow, from data preprocessing to model deployment. **Analyze** how major Python libraries can be integrated across different stages of the predictive modeling workflow in this scenario.	**3**	
2.	**Consider the following dataset:** 12, 7, 25, 18, 5, 40, 22, 9, 14, 30, 11, 6, 19, 8, 52			
	a)	Using the given dataset, determine the five-number summary and construct an appropriate boxplot. **Evaluate** the distribution of the data by interpreting the boxplot and **justify** whether any outliers exist, supporting your conclusion with logical reasoning.	**5**	**CLO-2** ***Level-5***
	b)	Suppose a data analyst ignores the detected outlier. **Critically evaluate and conclude** how this decision may affect data analysis or machine learning model performance.	**3**	

3.	Consider the following documents: • Doc1: *Bangladesh is an independent country.* • Doc2: *Bangladesh became independent with a liberation war.* • Doc3: *Myanmar became independent this year.* **Evaluate** the most appropriate algorithm for measuring document similarity among the given documents. Using the selected algorithm, assess the similarity between the document pairs (Doc1–Doc2, Doc1–Doc3, and Doc2–Doc3) and **justify** your results.	5	**CLO-2** ***Level-5***
4.	Consider the following text dataset: **Develop** naïve bayes algorithm to determine the class for the given test data.	5	**CLO-3** ***Level-6***

	Comments	*Class*
Training	*just playing boring.*	*Negative*
	entirely predictable and lacks energy.	*Negative*
	no surprises and very few laugh.	*Negative*
	very powerful, A+.	*Positive*
	the most fun film of the summer	*Positive*
	predictable with fun	*Neutral*
Test	*Predictable, recommended to all*	*?*