ourdiu.com | Shared by MEHEDI HASAN TUHIN

Daffodil International University

Faculty of Science & Information Technology

Department of Software Engineering

Midterm Examination, Summer 2026

Course Code: DS331; Course Title: Introduction to Data Science and Data Management & Analysis

Sections & Teachers: 41 & (A,B,C) MSA, (D,E) IAT, (G,H) SSI

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	In the modern world, almost every activity produces data. When people use smartphones, search on Google, watch videos on YouTube, buy products online, use mobile banking, visit hospitals, attend online classes, or travel using ride-sharing apps, they generate a huge amount of data. This data may include customer preferences, health records, transaction history, location information, learning patterns, product reviews, images, videos, and sensor readings. Many experts say, "**Data is the new oil.**" Oil became valuable because it powered industries, transportation, machines, and economic growth. In the same way, data has become one of the most valuable resources in the digital age. At the same time, technologies such as **artificial intelligence**, machine learning, automation, robotics, smart healthcare, autonomous vehicles, fraud detection systems, recommendation systems, and smart cities are growing rapidly. These technologies do not work magically; they need large amounts of meaningful data to learn, predict, and make intelligent decisions. However, raw oil is not useful until it is refined. Similarly, raw data is not very useful unless it is prepared into meaningful knowledge. Raw data must be prepared into meaningful knowledge before it can be used by artificial intelligence and machine learning systems. During this data preparation process, categorical data such as product type, payment method, customer location, disease type, or learning platform may need to be converted into numerical form.		***CLO-1 Level-2***
	a) Based on the above scenario, **explain** why data science is considered a foundation for future technologies such as artificial intelligence, machine learning, and automation.	[Marks 4]	
	b) From the given scenario, **differentiate** between Data Science, Data Analytics, and Data Management, highlighting their contributions to transforming raw data into meaningful knowledge for AI and machine learning systems.	[Marks 3]	

ourdiu.com | Shared by MEHEDI HASAN TUHIN

2 A university is trying to understand how much time students spend on online learning platforms each day. The collected data is given below: — *CLO-1 Level-2*

Name	Semester	Skill	Platform	Learning_Hours	Completion_%	Quiz_Score
Rahim	3	Python	Coursera	6	72	68
Karim	4	Data Science	Udemy	8	85	78
Nila	3	Web Development	YouTube	5	60	58
Sami	5	Machine Learning	Coursera		90	88
Rafi	4	Python	Udemy	7	78	74
Rafi	4	Python	Udemy	7	78	174
Rahim	3	Python	Coursera		72	68
Jui	4		Coursera	6	70	65
Rima	3	Python	YouTube	90	75	70
Nayeem	5	Data Science	edX	8	82	140

Review the given students' online learning dataset and prepare it for meaningful analysis by applying necessary data cleaning techniques. After cleaning the dataset, **visualise** the data using suitable visualizations and briefly interpret the major patterns, relationships — **[8 Marks]**

3 Turing Award winner Yann LeCun posted the following on X (Twitter) in September 2023: — **[5 Marks]** — *CLO-2 Level-5*

"Auto-Regressive LLMs can't plan (and can't really reason)."

— Yann LeCun (@ylecun), 13 September 2023

a) **Explain** autoregression in the context of LLMs, then draw and label a high-level block diagram of how a LLM works in general, making autoregression explicit in your diagram. — **[5 Marks]**

A plain LLM's knowledge is frozen at its training cut-off and has no access to private or organisation-specific documents. RAG addresses this by retrieving relevant content at query time rather than relying solely on what the model learned during training.

b) Create a high-level diagram of a Retrieval-Augmented Generation (RAG) and **explain** the complete flow from document corpus preparation (getting your own files ready) and indexing to answering a user's question with a grounded response.