

Question 1 (10 marks)

You are working on a credit-risk prediction model for a bank. The dataset contains the following columns:

- **Loan_ID** - Unique identifier for each loan
- **Applicant_Income** - Monthly income of the borrower
- **Loan_Purpose** - Category describing why the loan was taken (17 unique categories)
- **Default_Flag** - Target variable (1 = default, 0 = non-default)

Answer the following:

Q1.1 (2 marks)

Should Loan_ID be included as a feature in the model? Explain your reasoning.

Q1.2 (3 marks)

Loan_Purpose contains 17 categories with no ordinal meaning.

- (a) Should you use Ordinal Encoding? Why or why not?
- (b) Suggest a more suitable encoding method and briefly describe how it works.

Q1.3 (3 marks)

Applicant_Income is extremely right-skewed.

- (a) Name one transformation technique you can apply.
- (b) Explain why this transformation helps the model.

Q1.4 (2 marks)

Give one reason why feature scaling is important if using Logistic Regression or KNN for credit risk.

Question 2 (10 marks)

Dataset: Product reviews from an e-commerce website

Review_ID	Review_Text	Stars
R101	“Delivered quickly but packaging was damaged.”	3
R102	“Excellent quality! Worth every rupee.”	5
R103	“Product stopped working after two days.”	1
R104	“Good value for money but not very durable.”	4
R105	“Terrible experience, completely disappointed.”	1

Goal: Predict the **Stars** rating (1–5) from the review text.

Q2.1 (6 marks)

List **three feature engineering techniques** to convert Review_Text into numeric representations for modelling.

Explain *why* each technique is useful.

Q2.2 (4 marks)

Explain why **dimensionality reduction** is often necessary when working with text features.

Name one suitable technique and describe how it works.

Question 3 (10 marks)

Dataset: IoT sensor data from a smart building

Timestamp	Room_ID	Temperature	Humidity	Motion_Flag
2025-07-01 08:00	R12	26.3	48	1
2025-07-01 08:05	R12	26.8	47	1
2025-07-01 08:10	R12	27.1	46	0

2025-07-01 08:15	R12	27.4	45	0
2025-07-01 08:20	R12	27.9	44	1

Goal: Predict **Temperature** at the next timestamp.

Q3.1 (5 marks)

Suggest **three time-derived features** you can extract from Timestamp.

For each, explain *why* it is useful for temperature prediction.

Q3.2 (5 marks)

Humidity and Motion_Flag may carry more information than appears.

Propose **two feature engineering strategies** to make these features more predictive and explain how they help.

Question 4 (10 marks)

Answer the following:

Q4.1 (5 marks)

Explain what feature engineering involves when working with **audio (.wav)** files.

Q4.2 (5 marks)

Is it possible to train a model directly on raw audio waveforms without explicit feature extraction?

Explain when this is possible and when it is not.