# Q1

1.1 List two problems that regression algorithms can be used to solve [1 mark]

The weather forecast and house price forecast can use regression algorithms to solve.

1.2 List two problems that belong to classification problem [1 mark]

Spam detection and Image classification are classification problem.

1.3 List two problems that can be solved by using clustering algorithm(s) [1 mark]

Social Media User Categorization and News Article Auto-classification can be solved by clustering algorithm.

1.4 What is anomaly detection? [1 mark]

Anomaly detection is a technique used to identify rare items, events, or observations that deviate significantly from most data patterns. These anomalies may indicate critical incidents, such as fraud, defects, or security breaches.

1.5 Use an example to illustrate what the ML.NET pipeline is [2 marks]

ML.NET Pipeline for Detecting Microloan Fraud

An ML.NET pipeline is a structured workflow that automates the process of building, training, and deploying a machine learning model to detect suspicious microloan applications. Here’s how it works in the context of fraud detection:

1) Data Loading & Preparation

Input: A dataset containing loan applications with features like:

* Applicant’s credit score
* Loan amount requested
* Employment history
* Previous loan repayment behavior
* Device/IP location (to detect fake identities)

Handling Missing Data: The pipeline automatically fills or removes incomplete records.

2) Feature Engineering

The pipeline transforms raw data into meaningful fraud signals:

* Normalization: Scales numerical values (e.g., income, loan amount) to a standard range.
* Categorical Encoding: Converts text-based features (e.g., employment type) into numerical values.
* Derived Features: Creates new indicators, such as:
  + "Loan-to-Income Ratio" (high ratios suggest risk)
  + "Geographic Mismatch" (applicant’s IP ≠ stated address)

3) Model Training

Algorithm Selection: Common choices for fraud detection:

- Binary Classification (e.g., "Fraudulent" vs. "Legitimate")

- Anomaly Detection (if labeled fraud data is scarce)

Training Process: The pipeline feeds engineered features into the model to learn patterns, such as:

- Fraudsters often request small loans (to avoid scrutiny).

- Repeated applications from the same device but different identities.

4) Model Evaluation

Metrics: The pipeline calculates:

- Precision: % of flagged frauds that are real (avoid false alarms).

- Recall: % of actual frauds caught (minimize missed frauds).

Threshold Tuning: Adjusts sensitivity (e.g., flagging only high-confidence fraud cases).

5) Deployment & Real-Time Detection

The trained model integrates with loan approval systems to:

- Score new applications (e.g., "85% fraud probability").

- Trigger alerts for manual review.

Continuous Learning: The pipeline can retrain periodically with new fraud patterns.