

PREVIOUS YEAR SOLVED QUESTION BANK

— for —
BITS WILP MTECH AIML

MACHINE LEARNING

(UNTIL MID SEMESTER - EC2)

Rohith Raj Reganti | crackbitswilp.in

ML Previous Year Solved Question Papers

Until Mid Semester - For BITS WILP M.Tech

crackbitswilp.in

1st Edition | 2026

ML Previous Year Solved Question Papers
Until Mid Semester - For BITS WILP M.Tech

Copyright © 2026 crackbitswilp.in

All rights reserved. No part of this publication may be reproduced without prior written permission of the publisher.

Published by
Printed in India

Contents

1	Introduction & Machine Learning Workflows	1
2	Data Pre Processing	11
3	Linear Models For Regression	21
4	Linear Models For Classification	37
5	Model Evaluation	59
6	Decision Trees	75

Introduction & Machine Learning Workflows

..... crackbitswilp.in ☎ 8660242841

Q 1.1 2026 Jan / 2+2 Marks

Tom Mitchell defined Machine Learning as: “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E”.

Answer the following questions with respect to the above definition:

- Suppose you provide a learning algorithm with a lot of historical weather data to make a ML model that predicts whether the weather will be Sunny, Windy, or Rainy. In this setting, identify E, T, and P. Also name any two suitable performance measures.
- The quantity of rainfall in a day is measured in millimeters or inches. If a learning algorithm is used to forecast tomorrow’s rainfall, which learning technique would be suitable? Justify your choice and name any two suitable performance measures.

Solution

- Task (T):** Weather prediction (Sunny, Windy, Rainy), which is a classification task.
 - Experience (E):** Historical weather data used to train the model.
 - Performance Measure (P):** Correct classification of weather conditions.

Suitable performance measures:

- Accuracy
- Precision
- Recall
- F-score

- Suitable learning technique: **Regression**
 - Justification: The target variable (rainfall amount) is continuous.

Suitable performance measures:

- Mean Squared Error (MSE)
- R^2 (Coefficient of Determination)

Explanation

- Weather prediction involves assigning inputs to one of several predefined categories, which makes it a classification problem. Historical weather data represents the experience, and classification metrics measure predictive performance.
- Rainfall prediction outputs a numerical value, making regression the appropriate learning technique. Regression metrics evaluate how closely predicted rainfall values match actual values.

..... crackbitswilp.in ☎ 8660242841

Q 1.2 2025 July / 2 Marks

For the given problem statements, identify the most suitable type of learning (Supervised, Unsupervised, or Reinforcement) for each scenario. Justify your answer. [2]

- **Problem Statement 1:** A self-driving car must navigate urban environments, adjusting speed, avoiding obstacles, and obeying traffic rules while minimizing fuel consumption and travel time. The car improves performance through interaction with the environment and feedback from driving outcomes.
- **Problem Statement 2:** An online retail platform wants to predict whether a product review is positive, negative, or neutral using a large set of labeled customer reviews.
- **Problem Statement 3:** A geneticist wants to identify hidden patterns in DNA sequences using a dataset of unlabeled genetic codes to cluster similar structures.
- **Problem Statement 4:** An insurance company wants to predict the amount of insurance claim a customer might make based on historical data.

Solution

- **Problem Statement 1:** Reinforcement Learning
- **Problem Statement 2:** Supervised Learning – Classification
- **Problem Statement 3:** Unsupervised Learning – Clustering
- **Problem Statement 4:** Supervised Learning – Regression

Explanation

- Problem Statement 1 uses reinforcement learning because the agent (car) learns by interacting with the environment and receiving feedback in the form of rewards or penalties.
- Problem Statement 2 involves labeled data and categorical outputs, making supervised learning with classification the appropriate choice.
- Problem Statement 3 uses unlabeled data and aims to discover hidden patterns, which is characteristic of unsupervised learning through clustering.
- Problem Statement 4 predicts a continuous numerical value (claim amount), so supervised

learning using regression is suitable.

..... crackbitswilp.in ☎ 8660242841

Q 1.3 2025 June / 5 Marks

Consider that you are heading an AI company. ABZ Bank approached you to enhance their systems with AI integration in all possible aspects to improve the performance of the bank and also customer experience. Answer the following questions based on this requirement.

- i) Discuss any two possible areas which can be considered for building ML Models. [1]
- ii) Name the data (labelled / unlabelled) and the associated learning for the cases. [1]
- iii) Mention the data related requirements and challenges to start working with ML Models. [1.5]
- iv) Write detailed steps in designing ML systems for these requirements. [1.5]

Solution

i) Two Possible Areas for ML Models

1. **Document Processing & OCR for KYC Verification** – Automating the extraction and validation of information from customer identity and address proofs. *Ex: BAJAJ Finserve, SBI e-KYC process.*
2. **Credit Risk Prediction / Loan Default Prediction** – Predicting the likelihood of a customer repaying a loan. *Ex: Credit card approval by any bank.*

ii) Data Type and Associated Learning

1. **Document Processing & OCR:** Labelled Data (scanned documents with annotated text regions and the actual text). Learning: Supervised Learning (Computer Vision + NLP models).
2. **Credit Risk Prediction:** Labelled Data (historical loan data with repayment/default status). Learning: Supervised Learning (Classification) to classify whether a customer is risky or safe for loan sanction.

iii) Data Requirements and Challenges

Requirements:

- **OCR:** Large dataset of scanned KYC documents (IDs, utility bills, passports) with ground-truth text.
- **Credit Risk:** Historical loan application data, customer demographics, and repayment history.

Challenges:

- **OCR:** Poor quality/scanned images, handwritten text, variations in document formats.
- **Credit Risk:** Imbalanced data (fewer defaults vs. successful repayments), missing or incorrect customer details.
- **Privacy & Security:** Handling sensitive personal and financial data while complying with regulations.

iv) Detailed Steps in Designing ML Systems

1. **Requirement Analysis** – Define objectives (reduce KYC onboarding time, improve loan approval decisions).
2. **Data Collection & Preprocessing**
 - *OCR*: Collect diverse document scans, apply noise reduction, text region annotation.
 - *Credit Risk*: Gather loan histories, normalize income/credit score data, handle class imbalance.
3. **Feature Engineering**
 - *OCR*: Extract visual/text features, embeddings for names/addresses.
 - *Credit Risk*: Create features like income-to-loan ratio, repayment history, and credit utilization.
4. **Model Selection, Training & Evaluation**
 - *OCR*: CNN + Transformer-based OCR models; evaluated using Accuracy and Word Error Rate (WER).
 - *Credit Risk*: Classification models (Logistic Regression, Random Forest, XGBoost); evaluated using Precision, Recall, AUC-ROC.
5. **Deployment & Maintenance** – Embed OCR into KYC onboarding systems; integrate risk models with loan approval workflow. Regularly retrain with new data, monitor for drift, and ensure compliance with banking standards.

..... crackbitswilp.in ☎ 8660242841

Q 1.4 2025 Feb / 2 Marks

For the given problem statements, identify the most suitable type of learning (Supervised, Unsupervised, or Reinforcement) for each scenario. Justify your answer. [2]

- **Problem Statement 1:** A self-driving car must navigate through urban environments, adjusting speed, avoiding obstacles, and obeying traffic rules while minimizing fuel consumption and travel time. The car improves its performance over time through interactions with the environment and feedback from its driving outcomes.
- **Problem Statement 2:** An online retail platform wants to predict whether a product review is “positive”, “negative”, or “neutral” using a large set of previously labeled customer reviews. The goal is to automate sentiment analysis for improving customer feedback handling.
- **Problem Statement 3:** A geneticist wants to identify hidden patterns in DNA sequences by analyzing a dataset of unlabeled genetic codes. The goal is to cluster similar genetic structures together to identify potential links to diseases, without any prior classification.
- **Problem Statement 4:** An insurance company wants to predict the amount of insurance claim a customer might make in a year based on their age, driving history, vehicle type, and location. The dataset contains historical records of customers and their claim amounts.

Solution

- **Problem Statement 1: Reinforcement Learning** – The car learns by interacting with the environment and receiving feedback (rewards/penalties) based on its driving outcomes. There is no labeled dataset; the agent improves its policy through trial and error to maximize cumulative reward (minimize fuel consumption and travel time).
- **Problem Statement 2: Supervised Learning – Classification** – The platform has a large set of *labeled* customer reviews (positive, negative, neutral). The model learns a mapping from review text to sentiment class, making this a multi-class classification problem under supervised learning.
- **Problem Statement 3: Unsupervised Learning – Clustering** – The genetic dataset is *unlabeled* with no prior classification. The goal is to discover hidden patterns and group similar DNA sequences together, which is the core objective of unsupervised clustering (e.g., K-Means, Hierarchical Clustering).
- **Problem Statement 4: Supervised Learning – Regression** – The dataset contains historical records with a continuous target variable (claim amount). The model learns from labeled input-output pairs to predict a numeric value, making this a supervised regression problem.

..... crackbitswilp.in ☎ 8660242841

Q 1.5 2024 July / 2 Marks

For the given problem statements, identify the type of learning most suitable for developing the ML solution. Justify your stance. [2]

- **Problem Statement 1:** An autonomous warehouse uses robotic arms to pick and place items into bins. The robot must learn optimal strategies for handling items of varying sizes and shapes, ensuring minimal errors while maximizing speed and efficiency. The robot learns through repeated interactions with the environment, receiving feedback based on its actions.
- **Problem Statement 2:** A financial institution wants to classify incoming loan applications as either “Approved” or “Rejected” based on historical data containing labeled instances of approved and rejected applications. The goal is to automate the approval process to save time and reduce human intervention.
- **Problem Statement 3:** A researcher aims to group articles from a large, unlabeled dataset into clusters based on their content similarity, enabling the discovery of underlying topics and themes. The task does not involve any predefined categories or labels for the articles.
- **Problem Statement 4:** A healthcare provider wants to estimate a patient’s cholesterol level (in mg/dL) based on their age, weight, exercise habits, diet, and genetic information.

Solution

- **Problem Statement 1: Reinforcement Learning** – The robotic arm learns through repeated interactions with the environment and receives feedback (rewards/penalties) based on the outcomes of its actions. There is no labeled dataset; the agent improves its picking and placing