1. **Predict the price of house**

   Scenario: A real estate company wants to predict the price of a house based on square footage, number of bedrooms, and location.
   Q: Identify the problem type and outline the step-by-step logic to solve it.

   Define the Problem:

   - This is a **Regression** problem in machine learning. Since the goal is to predict the price of a house.

   Collect & Understand the Data:

   - **Input features(X):** Square footage, Number of bedrooms, Location
   - **Target variable(y):** House price
   - This is a **Regression** task.

   Data Preprocessing:

   - Handle missing values
   - Encode Categorical Variable (Location)

   Split the Data:

   - Divide into Training and Test set, this helps evaluate real-world performance.
   - Training set (70-80%)
   - Test set (20-30%)

   Choose the ML Model:

   - Use a regression model like Linear Regression or Decision Tree Regression or Random Forest Regressor, Gradient Boosting (XGBoost, LightGBM)

   Train the Model:

   - Fit the model on the training dataset.
   - The model learns a function: Price=f(SqFt,Bedrooms,Location)

   Evaluate the Model:

   Use regression metrics:

   - MAE (Mean Absolute Error) – easy to interpret
   - RMSE (Root Mean Squared Error) – penalizes large errors
   - R^2 Score – variance explained

`Make Predictions:

- Use the model to predict house prices for new data.

2. **Identifying Fraudulent Transactions**

Scenario: A bank wants to build a model to detect fraudulent transactions by analyzing customer spending behavior and transaction history.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Gather transaction records labeled as fraudulent or non-fraudulent.
- This is typically a binary Classification problem.

Collect & Understand the Data:

- You need features like Transaction amount, Date & Time, Transaction ID, Location.

Data preprocessing & cleaning:

- Handle missing values
- Encode categorical features
- Remove outliers, normalize transaction amounts

Split the Data:

- Divide into Training and Test set, this helps evaluate real-world performance.

Choose the ML Model:

- Use a regression model like Logistics Regression, Decision Tree, Random Forest, Gradient Boosting (XGBoost, LightGBM)

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use metrics like accuracy, precision, recall, and F1-score

Deploy Model:

- Implement real-time fraud detection.

3. **Grouping Customers Based on Spending Habits**

A supermarket wants to segment its customers based on their shopping patterns to provide personalized promotions.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Clustering

Collect & Understand the Data:

- You need features like customer purchase history, amount spent, and frequency of purchase

Data preprocessing & cleaning:

- Handle missing values
- Encode categorical features
- Remove outliers, normalize transaction amounts

Split the Data:

- Divide into Training and Test set, this helps evaluate real-world performance.

Choose the ML Model:

- Use a regression model like Logistics Regression, Decision Tree, Random Forest, Gradient Boosting (XGBoost, LightGBM)

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use metrics like accuracy, precision, recall, and F1-score

Deploy Model:

- Implement real-time fraud detection

4. Predicting Employee Salaries

Scenario: A company wants to estimate an employee's salary based on years of experience, job title, and education.

Define the Problem:

- This is a **Regression** problem in machine learning. Since the goal is to predict a continuous value (Salary)

Collect & Understand the Data:

- **Input features(X):**

  Years Of Experience - Numerical (Continuous/Discrete)

  Job Title - Categorical (Nominal)

  Education - Categorical (Ordinal - because there is a hierarchy, e.g., Master's > Bachelor's).

- **Target variable(y):**

  Salary (Numerical)

Data preprocessing & cleaning:

- Handle missing values
- Encode categorical features (Ex: Job title)

Split the Data:

- Divide into Training and Test set, this helps evaluate real-world performance.

Choose the ML Model:

- Use a regression model like Linear Regression, Random Forest

Train the Model:

- Fit the model on the training dataset.

Evaluate the Model:

Use regression metrics:

- MAE (Mean Absolute Error) – easy to interpret
- RMSE (Root Mean Squared Error) – penalizes large errors
- R^2 Score – variance explained

`Make Predictions:

- Use the model to predict the employee's salary for new data.

5. Detecting Spam Emails

An email provider wants to automatically classify incoming emails as spam or not spam based on their content and sender details.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Clustering

Collect & Understand the Data:

- You need features like customer purchase history, amount spent, and frequency of purchase

Data preprocessing & cleaning:

- Handle missing values
- Encode categorical features
- Remove outliers, normalize transaction amounts

Split the Data:

- Divide into Training and Test set, this helps evaluate real-world performance.

Choose the ML Model:

- Use a regression model like Logistics Regression, Decision Tree, Random Forest, Gradient Boosting (XGBoost, LightGBM)

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use metrics like accuracy, precision, recall, and F1-score

Deploy Model:

- Implement real-time fraud detection

6. Sentiment Analysis

A business wants to analyze customer reviews of its products and determine whether the sentiment is positive or negative.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Classification

Collect & Understand the Data:

- Input: Customer review text
- Output: Sentiment label (Positive or Negative)

Data preprocessing & cleaning:

- Convert text to lowercase.
- Remove punctuation, numbers, and special characters.
- Remove stop words (e.g., *is, the, and*).
- Handle misspellings or emojis if relevant.

Split the Data:

- Divide into Training and Test set.

Choose the ML Model:

- Use a regression model like Logistic Regression, Naive Bayes, Support Vector Machines (SVM), Neural networks (e.g., LSTM, Transformers)

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use metrics like accuracy, precision, recall, and F1-score to check the model performances.

Deploy Model:

- Deploy the model to classify new customer reviews as positive or negative.

7. Insurance Claim Prediction

An insurance company wants to predict whether a customer is likely to file a claim in the next year based on their driving history and demographics.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Supervised learning classification problem, specifically binary classification
- Predict whether a customer will file an insurance claim in the next year using past data

Collect & Understand the Data:

- Input: Vehicle details,cDriving history, Past claim history
- Output: Filed claim? (Yes/No)

Data preprocessing & cleaning:

- Handle missing values
- Remove duplicates
- Encode categorical variables (e.g., gender, location)
- Normalize or scale numerical features if needed
- Address class imbalance (e.g., claims are rare)

Split the Data:

- Training Set (70–80%)
- Test Set (20–30%)

Choose the ML Model:

- Use a regression model like Logistic Regression, Decision Tree, Random Forest, Gradient Boosting

Train the Model:

- Fit the model using the training data
- Optimize model parameters (hyperparameter tuning)

Evaluate the Model:

- Use metrics like accuracy, precision, recall, and F1-score and ROC-AUC (important for risk prediction) to check the model performances.

Deploy Model:

- Deploy the model to predict claims likelihood for new customers.

8. Sentiment Analysis

A streaming platform wants to recommend movies to users by grouping them based on their viewing preferences and watch history.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Clustering

Collect & Understand the Data:

- Gather user movie preferences, genres watched, and ratings

Preprocess Data:

- Convert categorical movie genres into numerical format.

Choose Clustering Algorithm:

- Use K-Means or Hierarchical Clustering.

Determine Optimal Clusters:

- Use the Elbow Method.

Train Model:

- Apply clustering algorithms to group users.

Analyze Clusters:

- Identify user categories (e.g., "Action Lovers," "Drama Fans").

Recommend Content:

- Suggest movies based on cluster preferences

9. Predicting Patient Recovery TIme:

A hospital wants to predict the recovery time of patients after surgery based on their age, medical history, and lifestyle habits.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Regression

Collect & Understand the Data:

- Gather historical recovery data with features like patient age, medical history, and lifestyle habits.

Data preprocessing & cleaning:

- Normalize medical features and handle missing values.

Split the Data:

- Divide into Training and Test set.

Choose the ML Model:

- Use a regression model like Random Forest Regression Linear Regression

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use RMSE (Root Mean Square Error) to check accuracy.

Deploy Model:

- Predict recovery time for new patients based on medical records.

10. Predicting Patient Recovery TIme:

A university wants to predict a student's final exam score based on study hours, attendance, and past academic performance.
Q: Identify the problem type and outline the step-by-step logic to solve it.

Define the Problem:

- Regression

Collect & Understand the Data:

- Gather historical student records with study hours, attendance, and exam scores.

Data preprocessing & cleaning:

- Handle missing values and standardize numerical features.

Split the Data:

- Divide into Training and Test set.

Choose the ML Model:

- Use a regression model like Linear Regression or Support Vector Regression.

Train the Model:

- Fit the model on the training dataset using labeled transaction data.

Evaluate the Model:

- Use metrics like RMSE (Root Mean Square Error) and R^2 score to check accuracy.

Make Predictions:

- Estimate exam scores for new students based on input features.