1. A real estate company wants to develop a system that predicts house prices based on square footage, number of bedrooms, and location.  
    **Q:** Identify the problem type and outline the step-by-step logic to solve it.

**Ramishahope Artificial Intelligence Pvt Ltd**

**36, Old Anandas, SG Arcade, Marudhamalai Main Road, Vadavalli, Coimbatore -641041.**

**+91 6385383227 |** [**www.hopelearning.net**](http://www.hopelearning.net/) **|** [**mdaravind@hopelearning.net**](mailto:mdaravind@hopelearning.net) **| 33AAMCR3722R1ZU**



**Problem Type:**

* **Regression** - because the goal is to predict a continuous numeric value (house price).

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather a dataset that includes input like square footage, number of bedrooms, location, and the corresponding house prices.
* **Explore and Analyze Data:** Understand the data with visualizations and statistics:
* **Check for missing values:** Identify correlations between variables
* **Preprocess the Data:** Handle missing values (e.g., imputation). Encode categorical variables like "location" (e.g., one-hot encoding or label encoding). Normalize or scale numerical input if needed.
* **Split Input and Output** **Variable:** Input (X): square footage, bedrooms, location Output (y): house price.
* **Split the Data into Training and Testing Sets:** Use something like an 80-20 or 70-30 split using train\_test\_split().
* **Build the Model:** Choose a regression model (e.g., Linear Regression, Decision Tree Regressor, Random Forest, etc.) and train it using the training data.
* **Evaluate the Model:** R² Score
* **Make Predictions:**Use the model to predict prices on new or unseen data.

1. 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.

**Problem Typ**e:

* **Classification** – because the goal is to classify transactions as either "fraudulent" or "not fraudulent"

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather a dataset that includes Transaction amount, Time of transaction, Customer ID, Location, Historical behavior, Fraud label (0 = not fraud, 1 = fraud)
* **Explore and Analyze Data:** Check for missing or inconsistent values.Explore distribution of input.Check class imbalance (fraud cases are often much fewer)
* **Preprocess the Data:** Handle missing values. Encode categorical variables (e.g., merchant type, location). Normalize or scale numerical input if required.
* **Split Input and Output Variable**: Input (X): All transaction and behavior data. Output (y): Fraud label (0 or 1).
* **Split the Data into Training and Testing Sets:** Use something like an 80-20 or 70-30 split using train\_test\_split().
* **Build the Model:** Choose a classification model (e.g., Linear Regression, Decision Tree Regressor, Random Forest, XGBoost etc.) and train it using the training data.
* **Evaluate the Model:** Accuracy, Precision, Recall, F1 Score, Confusion Matrix and ROC-AUC score.
* **Make Predictions:** Use the model to detect/predict if a transaction is fraudulent (1) or not (0)

1. 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.

**Problem Type:**

* Clustering – because the goal is to group customers based on their shopping patterns without predefined labels. This is an unsupervised learning problem.

**Step-by-Step Logic to Solve the Problem:**

**Collect Data:**

* **Gather customer-related data such as:** Customer ID, Transaction history: Recency, Frequency, Monetary value (RFM analysis), Shopping categories, Time and day of shopping, Location (optional), Behavioral patterns (e.g., loyalty, discount usage)
* **Explore and Analyze Data:** Check for missing or inconsistent values. Visualize patterns using plots (e.g., histograms, boxplots, heatmaps)
* **Preprocess the Dat**a: Handle missing values. Encode categorical variables (e.g., One-Hot Encoding). Normalize or scale numerical input (e.g.,StandardScaler).
* **Apply Clustering Algorithm:** KMeans, Agglomerative, DBSCAN, AffinityPropagation, etc. For KMeans, use the elbow method to determine the optimal number of clusters. Fit the model to your processed data
* **Evaluate and Visualize Clusters:** Use PCA to reduce dimensions for visualization. Plot clusters to interpret their separation. Use metrics like Silhouette Score to evaluate clustering quality.
* **Use Cluster Insights**: Analyze and label clusters (e.g., "Bargain Shoppers", "Loyal Customers", "High Spenders"). Use these insights to design personalized marketing strategies and promotional offers.

1. A company wants to estimate an employee’s salary based on their years of experience, job title, and education level.  
    **Q:** Identify the problem type and outline the step-by-step logic to solve it.

**Problem Type:**

* **Regression** – because the goal is to predict a continuous value (salary).

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Get employee data like years of experience, job title, education level, and salary.
* **Analyze Data:** Check for missing values and understand patterns in the data.
* **Preprocess Data**: Fill or remove missing values. Convert job title and education level into numbers (encoding).
* **Split Input and Output:** Input (X): experience, job title, education level. Output (y): salary
* **Split into Train and Test Data:** Use 80% for training, 20% for testing.
* **Build the Model:** Use a regression model like Linear Regression or Random Forest.
* **Train the Model:** Fit the model using the training data.
* **Evaluate the Model:** Use R² score
* **Make Predictions:** Predict salary for new employees based on their details.

1. 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.

**Problem Type:**

* **Classification** – because the goal is to classify emails as spam or not spam.

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather emails with labels (spam or not spam) along with their content and sender details.
* **Analyze Data:** Understand common words or patterns in spam and non-spam emails.
* **Preprocess the Data:**
  + Clean the text (remove punctuation, lowercase, etc.)
  + Convert text to numbers.
  + Encode sender details if needed
* **Split Input and Output:**
  + Input (X): Email content and sender info
  + Output (y): Spam (1) or Not Spam (0)
* **Split into Train and Test Sets:** Use train\_test\_split() to divide data (e.g., 80% train, 20% test)
* **Build and Train the Model:**Use a classification model like:
  + Naive Bayes (common for spam detection)
  + Logistic Regression
  + Random Forest, etc.
* **Test the Model:** Use the test set to check model performance using metrics like accuracy, precision, recall, F1 score
* **Make Predictions:** Use the trained model to predict if new incoming emails are spam or not.

1. 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.

**Problem Type:**

* **Classification –** because the goal is to determine whether a customer review is positive or negative, which is a binary classification task.

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather a dataset of customer product reviews labeled as positive or negative.
* **Analyze Data:** Understand common patterns and keywords that appear in positive vs. negative reviews.
* **Preprocess the Data:**
  + Clean the text (e.g., remove punctuation, convert to lowercase, remove stopwords)
  + Convert text to numbers.
* **Split Input and Output:** 
  + Input (X): Cleaned and encoded customer review text
  + Output (y): positive (1) or negative (0)
* **Split into Train and Test Sets:** Use train\_test\_split() to divide data (e.g., 80% train, 20% test).
* **Build and Train the Model:**

**Use a classification model like:**

* + Naive Bayes (common for spam detection)
  + Logistic Regression
  + Random Forest
  + Support Vector Machine (SVM)
* **Evaluate the Model:** Use the test set and evaluate the model using metrics like Accuracy, Precision, Recall, F1 Score.
* **Make Predictions:** Use the trained model to classify new customer reviews as positive or negative.

1. 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.

**Problem Type:**

* **classification** – because we are predicting Yes or No (will the customer file a claim or not).

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather a dataset of customer, Include driving history, age, gender, past claims, etc.
* **Clean the data:** Handle missing values, convert text to numbers (encoding), and scale if needed.
* **Split the data:** Divide into training and testing parts.
* **Train a model:** Use a classification algorithm (like Logistic Regression or Random Forest).
* **Test the model:** Check how well it works using accuracy or F1-score.
* **Improve if needed:** Tune settings or try different models.
* **Use the model:** Predict for new customers.

1. 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.

**Problem Type:**

* Clutering – Because we want to group users based on similar watching habits without labels.

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Gather user watch history, movie genres watched, ratings given, watch time, and other viewing behavior data**.**
* **Analyze Data:** Understand viewing patterns, genre preferences, and common behaviors across users.
* **Preprocess the Data:** 
  + Convert categorical data (e.g., genres) into numerical form using encoding.
  + Normalize numerical features (e.g., watch time, rating frequency).
  + Handle missing values if any.
* **Build User Profiles:** Combine each user's viewing behavior into a single feature vector representing their preferences.
* **Determine Optimal Number of Clusters:** Use the Elbow Method or Silhouette Score to choose the best number of user groups (clusters).
* **Apply Clustering Algorithm:**

**Use an unsupervised algorithm like:**

* + K-Means Clustering
  + Hierarchical Clustering
  + DBSCAN (for noise/outliers)
* **Analyze Clusters:** Interpret each group to understand user types (e.g., “Action Lovers,” “Rom-Com Fans,” “Documentary Viewers”).
* **Recommend Movies:** Suggest movies to a user based on popular content within their cluster.

1. 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.

**Problem Type:**

* **Regression** – Because the goal is to predict a continuous numerical value (recovery time in days/weeks).

**Step-by-Step Logic to Solve the Problem:**

* **Collect Data:** Get information like patient age, health history, lifestyle habits, and recovery time.
* **Preprocess Data:**
  + Handle missing values
  + Convert text to numbers (e.g., smoker: yes/no → 1/0)
  + Scale numeric values if needed
* **Split Input and Output:**
  + Input (X): Patient details
  + Output (y): Recovery time
* **Split into Train and Test Sets:** Use 80% for training, 20% for testing
* **Choose and Train a Model:** Use a regression model like Linear Regression or Random Forest.
* **Evaluate the Model:** Test the model using accuracy metrics R² score.
* **Make Predictions:** Use the trained model to predict recovery time for new patients.

1. 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.

**Problem Type:**

* **Regression**: Because the goal is to predict a numerical value (exam score).

**Step-by-Step Logic:**

* **Collect Data:** Gather data on each student’s study hours, attendance rate, previous grades, and final exam score.
* **Preprocess Data:** 
  + Handle missing values
  + Convert categorical data (if any) into numbers
  + Normalize/scale features (if needed)
* **Split the Data:** Divide into training set and test set (e.g., 80% training, 20% testing).
* **Choose a Regression Model:**

**Use models like**:

* + Linear Regression
  + Random Forest Regressor
  + Support Vector Regression (SVR)
* **Train the Model:** Use the training data to teach the model how the input features relate to the exam score.
* **Evaluate the Model:** Test the model using metrics like R² Score
* **Make Predictions:** Use the trained model to predict final exam scores for new or future students.