ML Scenario Based set1 Questions

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.

Logic:

1. **Collect the Data**
   - Collect historical data: *Square Footage, *Number of Bedrooms, *Location, *House Prices
2. **Domain Selection**
   - Task: **Machine Learning – Supervised Regression**
   - Objective: Predict house prices
3. **Data Preprocessing**
   - Handle categorical variables: **Location** is nominal → apply **One-Hot Encoding**
   - Handle missing values, if any
4. **Split Input and Output**
   - **Input Features (X):** Square Footage, Number of Bedrooms, Location
   - **Output (y):** House Prices
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Linear Regression**, **Decision Tree Regression**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Use **R^2 Score** or other regression metrics to assess performance
9. **Make Predictions**: Use the trained model to **predict house prices** for new data

2. 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.

Logic:

1. **Collect the Data**
   - Collect historical data: amount, location, transaction frequency, average spending, and unusual behavior detection, label = fraud/not fraud
2. **Domain Selection**
   - Task: **Machine Learning – Supervised Classification**
   - Objective: Fraud (1) / Not Fraud (0)
3. **Data Preprocessing**
   - Handle categorical variables: **Location** is nominal → apply **One-Hot Encoding**
   - Handle missing values, if any
4. **Split Input and Output**
   - **Input Features (X):** amount, location, transaction frequency, average spending, and unusual behavior detection
   - **Output (y):** label = fraud/not fraud

5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Logistic Regression**, **Random Forest**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Evaluate with metrics (Precision, Recall, F1-score, ROC-AUC).
9. **Make Predictions**: Predict fraud for new transactions

3. 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.

Logic:

1. **Collect Data** – Purchase history, amount spent, frequency of purchases.
2. **Domain Selection**: Unsupervised Learning – **Clustering** (Customer Segmentation)
3. **Preprocess Data** – Clean, handle missing values, normalize/spending scaling.
4. **Select Clustering Algorithm** – K-Means or Hierarchical.
5. **Find Optimal Clusters** – Use Elbow Method / Silhouette Score.
6. **Train Model** – Apply chosen algorithm to group customers.
7. **Analyze Clusters** – Identify high, medium, and low spenders.
8. **Apply Results** – Use clusters for personalized promotions and marketing strategies.

4. 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.

Logic:

1. **Collect the Data**
   - Collect historical data: *Education, * Years of Experience, * Employee's salary,
2. **Domain Selection**
   - Task: **Machine Learning – Supervised Regression**
   - Objective: Predict Employee's salary
3. **Data Preprocessing**: Education is nominal → apply **One-Hot Encoding,** Handle missing values, if any
4. **Split Input and Output**
   - **(X):** Education, Years of Experience
   - **Output (y):** Employee's salary
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Linear Regression**, **Decision Tree Regression**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Use **R^2 Score** or other regression metrics to assess performance
9. **Make Predictions**: Use the trained model to **predict Employee's salary** for new data

5. 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.

Logic:

1. **Collect the Data**
    - Collect historical data: Subject, Text, detection label = Spam/not Spam
2. **Domain Selection**
    - Task: **Machine Learning – Supervised Classification**
    - Objective: Spam (1) / Not Spam (0)
3. **Data Preprocessing**
    - Handle variables: Convert text into numbers **(TF-IDF).**
    - Handle missing values, if any
4. **Split Input(X) and Output(y)**: **X:** Subject, Text, **y:** label = Spam /not Spam
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Logistic Regression**, **Naive Bayes**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Evaluate with metrics (Precision, Recall, F1-score, ROC-AUC).
9. **Make Predictions**: Predict Spam /not Spam for new emails.

6. 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.

Logic:

1. **Collect the Data**
    - Collect historical data: customer reviews, detection label = Positive or Negative
2. **Domain Selection**
    - Task: **Machine Learning – Supervised Classification**
    - Objective: Positive (1) / Negative (0)
3. **Data Preprocessing**
    - Handle variables: Convert reviews into numbers **(TF-IDF or Word2Vec).**
    - Handle missing values, if any
4. **Split Input(X) and Output(y)**: **X:** Reviews **y:** label = Positive or Negative
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Logistic Regression**, **Naive Bayes**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Evaluate with metrics (Precision, Recall, F1-score, ROC-AUC).
9. **Make Predictions**: Predict Positive /not Negative for new Customer Reviews.

7. 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.

Logic:

1. **Collect the Data**
   - Collect historical data: Driving History, Demographics. label = Claim: yes or No
2. **Domain Selection**
   - Task: **Machine Learning – Supervised Classification**
   - Objective: Claim yes (1) / No (0)
3. **Data Preprocessing**
   - Handle variables: Convert text into numbers **(TF-IDF or Word2Vec).**
   - Handle missing values, if any
4. **Split Input(X) and Output(y)**: **X:** Subject, Text, **y:** label = Spam /not Spam
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Logistic Regression**, **Decision Tree**, **Naive Bayes**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Evaluate with metrics (Precision, Recall, F1-score, ROC-AUC).
9. **Make Predictions**: Predict Claim: yes or No for new Customer.

8. 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.

Logic:

1. **Collect Data** –user watch history (movies watched, ratings, time spent)
   Collect movie metadata (genre, actors, release year)
2. **Domain Selection**: **Unsupervised Learning** – **Clustering**
3. **Preprocess Data** – Clean, handle missing values, normalize scaling.
4. **Select Clustering Algorithm** – K-Means or Hierarchical or DBSCAN.
5. **Find Optimal Clusters** – Use Elbow Method / Silhouette Score.
6. **Train Model** – Apply chosen algorithm to group customers.
7. **Analyze Clusters** – Identify user segments (e.g., Action lovers, Comedy fans, Mixed viewers).
8. **Apply Results** – Recommend top movies from each user's cluster for personalization.

9. 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.

Logic:

1. **Collect the Data**
   - Collect historical data: *Age, * medical history, * lifestyle habits, *Recovery time.

2. **Domain Selection**
    - Task: **Machine Learning – Supervised Regression**
    - Objective: Predict Recovery time
3. **Data Preprocessing**: nominal → apply **One-Hot Encoding,** Ordinal – Label Encoding, Handle missing values, if any
4. **Split Input and Output**
    - **(X):** Age, medical history, lifestyle habits
    - **Output (y):** Recovery time
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Linear Regression**, **Random Forest Regression**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Use **R^2 Score** or other regression metrics to assess performance
9. **Make Predictions**: Use the trained model to **predict** recovery time of patients for new patients.

10. 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.

Logic:

1. **Collect the Data**
    - Collect historical data: * study hours, * attendance, * past academic performance, * final exam score.
2. **Domain Selection**
    - Task: **Machine Learning – Supervised Regression**
    - Objective: Predict final exam score
3. **Data Preprocessing**: nominal → apply **One-Hot Encoding,** Ordinal **–** apply **Label Encoding**, Handle missing values, if any
4. **Split Input and Output**
    - **(X):** study hours, attendance, past academic performance
    - **Output (y):** final exam score
5. **Split Dataset**: Divide into **Training (70%)** and **Testing (30%)** sets
6. **Choose Algorithm**: Examples: **Linear Regression**, **Random Forest Regression**
7. **Train the Model**: Fit the chosen model on the **training dataset**
8. **Evaluate the Model**: Use **R^2 Score** or other regression metrics to assess performance
9. **Make Predictions**: Use the trained model to **predict** final exam score for new student.