ML Scenario Question - Answers

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:

    Problem Identification:

    > **Stage 1 : Domain selection – Machine learning**
    > **Stage 2: Learning selection – Supervised learning**
    > **Stage 3: Regression/ Classification - Regression**

    A: House price prediction

    - **Requirements are clearly specified**
    - **Output variable also present**
    - **As the data is labelled this is supervised learning. Since prediction is numerical values it is a regression problem.**
    - ➢ **Data collection – importing the csv dataset into pandas,**
    - ➢ **Data preprocessing – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,**
    - ➢ **Input output split , train and test split**
    - ➢ **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
    - ➢ **Evaluation – Use R2 score to evaluate the accuracy and performance**
    - ➢ **Prediction -** Test the model on 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:

    Problem Identification:

    > **Stage 1 : Domain selection – Machine learning**
    > **Stage 2: Learning selection – Supervised learning**
    > **Stage 3: Regression/ Classification - Classification**

    A: Credit Card prediction

    - **Requirements are clearly specified**
    - **Output variable also present**
    - **As the data is labelled this is supervised learning. Since prediction is categorical values it is a Classification problem.**
    - ➢ **Data collection – importing the csv dataset into pandas,**
    - ➢ **Data preprocessing – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,**
    - ➢ **Input output split , train and test split**
    - ➢ **Feature selection**

- ➢ **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
- ➢ **Evaluation – Use Confusion Matrix to evaluate the accuracy and performance**
- ➢ **Prediction -** Test the model on new data.

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.

   A: Customer Segmentation (Clustering – Unsupervised Learning)

   Logic:

   Problem Identification:
   - **Stage 1: Domain Selection – Machine Learning**
   - **Stage 2: Learning Selection – Unsupervised learning**
   - **Stage 3: Regression / Classification / Clustering – Clustering**

   A: Customer Segmentation
   - **Requirements are clearly specified**
   - **Output variable is not present**
   - **As the data is unlabelled, this is an unsupervised learning problem**
   - **Since the goal is to group similar customers based on shopping behavior, it is a clustering problem**
   - ➢ **Data collection – Import customer shopping data (CSV) into pandas**
   - ➢ **Data preprocessing – Handle missing values, encode categorical data, and scale features**
   - ➢ **Feature selection – Select relevant shopping behavior attributes**
   - ➢ **Model creation – Apply clustering algorithms like K-Means**
   - ➢ **Evaluation – Use Elbow Method or Silhouette Score to validate clusters**
   - ➢ **Prediction – Segment customers and use clusters for personalized promotions**

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.

   A: supervised learning regression

   Logic:

   Problem Identification:
   - **Stage 1 : Domain selection – Machine learning**
   - **Stage 2: Learning selection – Supervised learning**
   - **Stage 3: Regression/ Classification - Regression**

   A: Employee Salary prediction
   - **Requirements are clearly specified**
   - **Output variable also present**
   - **As the data is labelled this is supervised learning. Since prediction is numerical values it is a regression problem.**
   - ➢ **Data collection – importing the csv dataset into pandas,**

- ➢ **Data preprocessing – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,**
- ➢ **Input output split, train and test split**
- ➢ **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
- ➢ **Evaluation – Use R2 score to evaluate the accuracy and performance**
- ➢ **Prediction -** Test the model on 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.

   A: supervised learning classification

   Problem Identification:

           Stage 1 : Domain selection – Machine learning

           Stage 2: Learning selection – Supervised learning

           Stage 3: Regression/ Classification - Classification

   <mark>**A: Spam Classification prediction**</mark>
   - **Requirements are clearly specified**
   - **Output variable also present**
   - **As the data is labelled this is supervised learning. Since prediction is categorical values it is a Classification problem.**
   - ➢ **Data collection – importing the csv dataset into pandas,**
   - ➢ **Data preprocessing – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,**
   - ➢ **Input output split , train and test split**
   - ➢ **Feature selection**
   - ➢ **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
   - ➢ **Evaluation – Use Confusion Matrix to evaluate the accuracy and performance**
   - ➢ **Prediction -** Test the model on new data.

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.

   A: Natural Language Processing

   Logic:

   Problem Identification:

           Stage 1 : Domain selection – Natural Language Processing

           Stage 2: Learning selection – Supervised learning

           Stage 3: Regression/ Classification - Classification

   <mark>**A: Sentiment Analysis**</mark>
   - **Requirements are clearly specified**
   - **Output variable also present**

- As the data is text based it is NLP. And it has labelled data this is supervised learning. Since prediction is categorical values it is a Classification problem.
- ➢ **Data collection – importing the csv dataset into pandas,**
- ➢ **Data preprocessing – preprocessing the data –**
    - o **Removing all punctuations, hastags, urls, emojis, Text tokenization,**
    - o **converting text into vectors using count vectorizer and tf-idf vectorizer**
- ➢ **Input output split, train and test split**
- ➢ **Model creation using ML based algorithm Naïve bayes is best for NLP.**
- ➢ **Evaluation – Use Confusion Matrix to evaluate the accuracy and performance**
- ➢ **Prediction -** Test the model on new data.

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.

   A: Classification

   Logic:

   Problem Identification:
   - o **Stage 1: Domain Selection – Machine Learning**
   - o **Stage 2: Learning Selection – Supervised learning**
   - o **Stage 3: Regression / Classification – Classification**

   A: Insurance Claim Prediction
   - o **Requirements are clearly specified**
   - o **Output variable is present (Claim: Yes / No)**
   - o **As the data is labelled, it is supervised learning**
   - o **Since the output is categorical, it is a classification problem**
   - ➢ **Data collection – Import customer driving and demographic data into pandas**
   - ➢ **Data preprocessing – Handle missing values and encode categorical variables**
   - ➢ **Input-output split – Separate features and target, then perform train-test split**
   - ➢ **Model creation – Apply classification algorithms like Logistic Regression or Random Forest**
   - ➢ **Evaluation – Use Confusion Matrix, Accuracy, Precision, and Recall**
   - ➢ **Prediction – Predict whether a customer will file a claim in the next year**

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.

   A: Recommendation System (Clustering – Unsupervised Learning)

   Logic:

   Problem Identification:
   - • **Stage 1: Domain Selection –** Machine Learning
   - • **Stage 2: Learning Selection –** Unsupervised learning
   - • **Stage 3: Regression / Classification / Clustering –** Clustering

   A: Movie Recommendation
   - • Requirements are clearly specified
   - • Output variable is not present
   - • As the data is unlabelled, it is unsupervised learning

- Since users are grouped based on viewing behavior, it is a **clustering problem**

Step-by-Step Logic:

- ➢ **Data collection** – Import user watch history and preference data into pandas
- ➢ **Data preprocessing** – Clean data, encode categorical features, and normalize values
- ➢ **Feature selection** – Select relevant viewing behavior attributes
- ➢ **Model creation** – Apply clustering algorithms like K-Means or Hierarchical Clustering
- ➢ **Evaluation** – Use Elbow Method or Silhouette Score
- ➢ **Recommendation** – Recommend movies based on the user's cluster

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.

A: Regression

Logic:

Problem Identification:

        Stage 1 : Domain selection – Machine learning

        Stage 2: Learning selection – Supervised learning

        Stage 3: Regression/ Classification - Regression

A: Healthcare Patient recovery prediction

- **Requirements are clearly specified**
- **Output variable also present**
- **As the data is labelled this is supervised learning. Since prediction is numerical values it is a regression problem.**

- ➢ **Data collection – importing the csv dataset into pandas,**
- ➢ **Data preprocessing – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,**
- ➢ **Input output split , train and test split**
- ➢ **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
- ➢ **Evaluation – Use R2 score to evaluate the accuracy and performance**
- ➢ **Prediction -** Test the model on new data.

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:

Problem Identification:

        Stage 1 : Domain selection – Machine learning

        Stage 2: Learning selection – Supervised learning

        Stage 3: Regression/ Classification - Regression

A: Exam Score prediction

- **Requirements are clearly specified**
- **Output variable also present**
- **As the data is labelled this is supervised learning. Since prediction is numerical values it is a regression problem.**

- **Data collection** – importing the csv dataset into pandas,
- **Data preprocessing** – preprocessing the data – removing null, na values, converting categorical data into numbers using one hot or label encoding,
- **Input output split , train and test split**
- **Model creation apply grid search cv for experimenting with different hyper tuning parameter**
- **Evaluation** – **Use R2 score to evaluate the accuracy and performance**
- **Prediction -** Test the model on new data.