Scenario Based - Set Qn 1

1) Predicting House Prices

Scenario: A real estate company wants to predict the price of a house based on square footage, number of bedrooms, and location.

a. Identify the problem type: Regression

b. Step-by-step logic:

1. **Collect Data** – Gather historical data with features like square footage, number of bedrooms, and location.
2. **Preprocess Data** – Handle missing values, encode categorical variables (e.g., location).
3. **Split Dataset** – Divide the dataset into training and testing sets.
4. **Choose Algorithm** – Use a regression model like Linear Regression or Decision Tree Regression.
5. **Train the Model** – Fit the model on the training dataset.
6. **Evaluate Performance** – Use metrics like R^2 score.
7. **Make Predictions** – Use the model to predict house prices for new data.

2) Identifying Fraudulent Transactions

Scenario: A bank wants to detect whether a transaction is fraudulent or not based on transaction history and customer behavior.

a. Identify the problem type: Classification

b. Step-by-step logic:

1. **Collect Data** – Gather transaction records labeled as fraudulent or non-fraudulent.
2. **Preprocess Data** – Remove outliers, normalize transaction amounts, and encode categorical features.
3. **Feature Engineering** – Create features like transaction frequency and unusual behavior detection.
4. **Split Dataset** – Divide data into training and testing sets.
5. **Choose Algorithm** – Use classification models like Logistic Regression, Random Forest, or Neural Networks.
6. **Train the Model** – Fit the model using labeled transaction data.

7. **Evaluate Performance** – Use metrics like accuracy, precision, recall, and F1-score.
8. **Deploy Model** – Implement real-time fraud detection.

3) Grouping Customers Based on Spending Habits

Scenario: A supermarket wants to segment customers into different groups based on their shopping patterns.

a. Identify the problem type: Clustering

b. Step-by-step logic:

1. **Collect Data** – Gather customer purchase history, amount spent, and frequency of purchases.
2. **Preprocess Data** – Normalize data (e.g., scale spending amounts to avoid bias).
3. **Choose Clustering Algorithm** – Use K-Means, DBSCAN, or Hierarchical Clustering.
4. **Determine Optimal Clusters** – Use the Elbow Method to find the best number of clusters.
5. **Train Model** – Apply clustering algorithm to group customers.
6. **Analyze Clusters** – Interpret results to identify high-spending, medium-spending, and low-spending customer groups.

4) Predicting Employee Salaries

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

a. Identify the problem type: Regression

b. Step-by-step logic:

1. **Collect Data** – Gather employee records with years of experience, education, and salary.
2. **Preprocess Data** – Handle missing values and encode categorical variables (e.g., job title).
3. **Split Dataset** – Separate data into training and testing sets.
4. **Choose Algorithm** – Use Linear Regression or Random Forest Regression.
5. **Train the Model** – Fit the model on training data.
6. **Evaluate Model** – Use R^2 score for accuracy measurement.
7. **Make Predictions** – Predict salary based on new employee data.

5) Detecting Spam Emails

Scenario: An email provider wants to classify emails as either spam or not spam based on content and sender details.

a. Identify the problem type: Classification

b. Step-by-step logic:

1. **Collect Data** – Use datasets of spam and non-spam emails.
2. **Preprocess Data** – Convert email text to numerical format using TF-IDF or word embeddings.
3. **Split Dataset** – Divide data into training and testing sets.
4. **Choose Algorithm** – Use Naive Bayes, Support Vector Machines, or Neural Networks.
5. **Train the Model** – Fit the model using labeled email data.
6. **Evaluate Model** – Measure accuracy using Precision, Recall, and F1-score.
7. **Deploy Model** – Automatically classify incoming emails as spam or not spam.

6) Customer Reviews Sentiment Analysis

Scenario: A company wants to determine whether customer reviews about a product are positive or negative based on review text.

a. Identify the problem type: Classification

b. Step-by-step logic:

1. **Collect Data** – Gather labeled customer reviews (positive/negative).
2. **Preprocess Text Data** – Remove stopwords, punctuation, and tokenize words.
3. **Convert Text into Features** – Use TF-IDF or Word2Vec to convert text into numerical format.
4. **Split Dataset** – Train-test split.
5. **Choose Algorithm** – Use Logistic Regression, Naive Bayes, or Transformers (BERT).
6. **Train Model** – Fit the model on the training dataset.
7. **Evaluate Model** – Use accuracy and F1-score to assess model performance.
8. **Make Predictions** – Classify new customer reviews as positive or negative.

7) Predicting Car Insurance Claims

Scenario: An insurance company wants to predict whether a policyholder will file a claim in the next year.

a. Identify the problem type: Classification

b. Step-by-step logic:

1. **Collect Data** – Gather past claim history, driving behavior, and customer demographics.
2. **Preprocess Data** – Handle missing values and encode categorical features.
3. **Split Dataset** – Divide data into training and testing sets.
4. **Choose Algorithm** – Use Logistic Regression, Decision Tree, or Neural Networks.
5. **Train the Model** – Fit the model using past claims data.
6. **Evaluate Model** – Use Precision-Recall, AUC-ROC score.
7. **Deploy Model** – Predict claims likelihood for new customers.

8) Recommending Movies Based

Scenario: A streaming platform wants to group users into categories based on their movie preferences and recommend similar content.

a. Identify the problem type: Clustering

b. Step-by-step logic:

1. **Collect Data** – Gather user movie preferences, genres watched, and ratings.
2. **Preprocess Data** – Convert categorical movie genres into numerical format.
3. **Choose Clustering Algorithm** – Use K-Means or Hierarchical Clustering.
4. **Determine Optimal Clusters** – Use the Elbow Method.
5. **Train Model** – Apply clustering algorithm to group users.
6. **Analyze Clusters** – Identify user categories (e.g., "Action Lovers," "Drama Fans").
7. **Recommend Content** – Suggest movies based on cluster preferences

9) Predicting Patient Recovery Time

Scenario: A hospital wants to predict how long it will take for a patient to recover from surgery based on age, medical history, and lifestyle.

a. Identify the problem type: Regression

b. Step-by-step logic:

1. **Collect Data** – Gather historical recovery data with features like patient age, medical history, and lifestyle habits.
2. **Preprocess Data** – Normalize medical features and handle missing values.
3. **Choose Regression Algorithm** – Use Random Forest Regression or Linear Regression.
4. **Train Model** – Fit the model on training data.
5. **Evaluate Model** – Use R^2 score to check accuracy.
6. **Make Predictions** – Predict recovery time for new patients based on medical records.

10) Predicting Student Exam Scores

Scenario: A university wants to predict a student's exam score based on study hours, past performance, and attendance.

a. Identify the problem type: Regression

b. Step-by-step logic:

1. **Collect Data** – Gather historical student records with study hours, attendance, and exam scores.
2. **Preprocess Data** – Handle missing values and standardize numerical features.
3. **Split Dataset** – Divide data into training and testing sets.
4. **Choose Algorithm** – Use Linear Regression or Support Vector Regression.
5. **Train the Model** – Fit the model on training data.
6. **Evaluate Performance** – Use metrics like R^2 score.
7. **Make Predictions** – Estimate exam scores for new students based on input features