Scenario Based - Qns Set 2

1) Predicting Loan Default

Scenario: A bank wants to predict whether a loan applicant will default based on credit score, income, and past loan history.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather customer financial history, credit scores, and loan repayment records.
- **Preprocess Data** – Handle missing values, normalize numerical features, and encode categorical variables.
- **Split Dataset** – Divide the dataset into training and testing sets.
- **Choose Algorithm** – Use Logistic Regression, Decision Trees, or Random Forest.
- **Train Model** – Fit the model using labeled loan default data.
- **Evaluate Performance** – Use AUC-ROC, Precision, Recall, and F1-score.
- **Make Predictions** – Predict loan default for new applicants.

2) Forecasting Demand for a Retail Store

Scenario: A retail store wants to predict the demand for different products to optimize inventory levels.

a. Identify the problem type: Regression

b. Step-by-step logic:

- **Collect Data** – Gather past sales data, seasonal trends, and product demand.
- **Preprocess Data** – Handle missing values, normalize numerical data, and remove outliers.
- **Split Dataset** – Divide the data into training and testing sets.
- **Choose Algorithm** – Use Linear Regression, Random Forest Regression, or XGBoost.
- **Train Model** – Fit the model using historical demand data.
- **Evaluate Performance** – Use R^2 score for evaluation.
- **Make Predictions** – Forecast demand for upcoming sales periods.

3) Detecting Defective Products in Manufacturing

Scenario: A factory wants to detect whether a manufactured product is defective based on sensor readings and quality control data.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather sensor readings, production details, and defect labels.

- **Preprocess Data** – Handle missing values, normalize numerical values, and encode categorical features.
- **Split Dataset** – Divide the data into training and testing sets.
- **Choose Algorithm** – Use Decision Trees, Support Vector Machines, or Neural Networks.
- **Train Model** – Fit the model using labeled defect data.
- **Evaluate Performance** – Use accuracy, precision, recall, and F1-score.
- **Deploy Model** – Detect defective products in real time.

4) Classifying Medical Diagnoses

Scenario: A healthcare provider wants to classify patient symptoms into different disease categories.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather patient records with symptoms and diagnoses.
- **Preprocess Data** – Handle missing values, normalize medical test results, and encode categorical features.
- **Split Dataset** – Train-test split.
- **Choose Algorithm** – Use Random Forest, Naive Bayes, or Gradient Boosting.
- **Train Model** – Fit the model using labeled medical data.
- **Evaluate Model** – Use accuracy, confusion matrix, and F1-score.
- **Make Predictions** – Predict disease category based on patient symptoms.

5) Identifying Fake Online Reviews

Scenario: An e-commerce company wants to detect fake reviews posted by bots or fraudsters.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather a dataset of real and fake reviews.
- **Preprocess Data** – Tokenize text, remove stopwords, and vectorize using TF-IDF.
- **Feature Engineering** – Identify suspicious patterns like repetitive words, unnatural phrasing, and review frequency.
- **Split Dataset** – Divide data into training and testing sets.
- **Choose Algorithm** – Use Naive Bayes, Logistic Regression, or Transformer models.
- **Train Model** – Fit the model on labeled review data.
- **Evaluate Performance** – Use accuracy, F1-score, and confusion matrix.
- **Make Predictions** – Detect fake reviews in real-time.

6) Predicting Stock Market Trends

Scenario: A financial firm wants to predict stock price movement based on historical price data and market indicators.

a. Identify the problem type: Regression

b. Step-by-step logic:

- **Collect Data** – Gather historical stock prices, trading volumes, and economic indicators.
- **Preprocess Data** – Handle missing values, normalize price changes, and engineer features like moving averages.
- **Split Dataset** – Train-test split.
- **Choose Algorithm** – Use Random Forest Regression, LSTMs, or Gradient Boosting.
- **Train Model** – Fit the model on historical stock data.
- **Evaluate Performance** – Use RMSE and directional accuracy.
- **Make Predictions** – Forecast future stock price movements.

7) Detecting Fake Social Media Accounts

Scenario: A social media platform wants to identify and remove fake user accounts.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather account details, activity logs, and engagement patterns.
- **Preprocess Data** – Handle missing values, engineer features like average post frequency and follower ratio.
- **Split Dataset** – Divide into training and testing sets.
- **Choose Algorithm** – Use Random Forest, Support Vector Machines, or XGBoost.
- **Train Model** – Fit the model using labeled real and fake account data.
- **Evaluate Performance** – Use Precision, Recall, and F1-score.
- **Make Predictions** – Identify and flag fake accounts.

8) Optimizing Ad Targeting for Online Marketing

Scenario: A digital marketing company wants to show the most relevant ads to users based on their browsing behavior.

a. Identify the problem type: Clustering

b. Step-by-step logic:

- **Collect Data** – Gather user click behavior, browsing history, and demographic information.
- **Preprocess Data** – Convert categorical features into numerical format, handle missing data.
- **Choose Algorithm** – Use K-Means or Hierarchical Clustering.
- **Determine Optimal Clusters** – Use the Elbow Method.

- **Train Model** – Apply clustering algorithm to segment users.
- **Analyze Clusters** – Identify user groups (e.g., "Tech Enthusiasts," "Fashion Lovers").
- **Optimize Ads** – Deliver targeted ads based on cluster preferences.

9) Classifying Land Cover in Satellite Images

Scenario: A geospatial research team wants to classify different land types (forest, water, urban) using satellite images.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Use satellite images labeled with land types.
- **Preprocess Data** – Normalize pixel values, remove noise, and extract image features.
- **Split Dataset** – Divide into training and testing sets.
- **Choose Algorithm** – Use Decision Trees, Support Vector Machines, or CNN-based models.
- **Train Model** – Fit the model on labeled satellite images.
- **Evaluate Performance** – Use accuracy and confusion matrix.
- **Make Predictions** – Classify new satellite images into land cover types.

10) Predicting Customer Churn for a Subscription Service

Scenario: A streaming service wants to predict which users are likely to cancel their subscriptions.

a. Identify the problem type: Classification

b. Step-by-step logic:

- **Collect Data** – Gather user engagement data, subscription history, and interaction logs.
- **Preprocess Data** – Handle missing values and encode categorical variables.
- **Feature Engineering** – Create features like average watch time and last login frequency.
- **Split Dataset** – Train-test split.
- **Choose Algorithm** – Use Logistic Regression, Random Forest, or Gradient Boosting.
- **Train Model** – Fit the model using past churn data.
- **Evaluate Performance** – Use AUC-ROC, Precision, and Recall.
- **Make Predictions** – Identify customers likely to churn and apply retention strategies.