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.
   Answer:
   problem type = regression problem (as the house price is continuous value)
   step-by-step logic:
   - Load the realEstate_houseprices dataset
   - Convert the categorical data (location column) into numerical data using pd.get_dummies
   - Split input (square_footage, number_of_bedrooms, location) and output (house_price)
   - Split train and test dataset
   - Data preprocessing for X_train using StandardScaler – as the square footage, no of bedrooms has large-scale difference
   - Model creation using regression algorithms (linear regression, SVR, decision tree, random forest, boosting etc…)
   - Test the model prediction
   - Evaluate the model using r2_score
   - Finalize the best model (by comparing all algorithm performance for the dataset)

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.
   Answer:
   Problem Type: Clustering Problem
   Step by Step logic:
   - Load the bank_customer dataset
   - Select the important features for X input data - (spending behavior, transaction history)
   - Data preprocessing – Standardization
   - Model creation & prediction of Y_labels [cluster group] – (using clustering algorithms: K means, agglomerative, DBSCAN, optics, birch etc.)
   - Create a new dataset with the cluster group & save it as ‘cluster.csv’ file
   - Visualize the clusters using seaborn.lmplot – (for visual inspection of clusters)
   - Cluster model evaluation using cluster evaluation metrics
   - Finalize the best cluster model for the problem statement – (by analyzing the cluster outputs, evaluation metric, domain knowledge and visual cluster plot)

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.
   Answer:
   Problem Type: Clustering Problem

Step by Step logic:

- Load the supermarket_customer dataset
- Select the important features for X input data - (shopping patterns:- spending behavior, income, savings, age)
- Data preprocessing – Standardization
- Model creation & prediction of Y_labels [cluster group] – (using clustering algorithms: K means, agglomerative, DBSCAN, optics, birch etc.)
- Create a new dataset with the cluster group & save it as 'cluster.csv' file
- Visualize the clusters using seaborn.lmplot – (for visual inspection of clusters)
- Cluster model evaluation using cluster evaluation metrics
- Finalize the best cluster model for the problem statement – (by analyzing the cluster outputs, evaluation metric, domain knowledge and visual cluster plot)

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.
   Answer:
   Problem Type: Regression Problem
   Step by Step logic:
   - Load the company_emplyees dataset
   - Convert the categorical data (job_title, educational_level column) into numerical data using pd.get_dummies
   - Split input (years_of_experience, job_title, education_level) and output (employee_salary)
   - Split train and test dataset
   - Model creation using regression algorithms (linear regression, SVR, decision tree, random forest, boosting etc…)
   - Test the model prediction
   - Evaluate the model using r2_score

   Finalize the best model (by comparing all algorithm performance for the dataset)

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.
   Answer:
   Problem Type: Classification Problem
   Step by Step logic:
   - Load the email_dataset
   - Split input (message_content, sender_details) and output (email_classify)
   - Split train & test dataset
   - Model creation using classification algorithms.

Must try algorithm: Naïve bayes – multinomial (good for detecting the spam or not spam emails as it works by computing the text occurrence count)

- Test the model
- Model evaluation using confusion matrix & classification report
- Finalize the best model for the Email classification of spam or not based on the performance analyses from confusion matrix & classification report of all algorithms

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.
   Answer:
   Problem Type: Classification Problem
   Step by Step logic:
   - Load the customer_reviews dataset
   - Split input (customer_reviews) and output (sentiment- positive/negative)
   - Split train and test dataset
   - Model creation using classification algorithms.

Must try algorithm: Naïve bayes – multinomial (good for detecting the positive & negative customer sentiment as it works by computing the text occurrence count)

- Test the model
- Model evaluation using confusion matrix & classification report
- Finalize the best model for the customer review sentiment analysis on the performance analyses from confusion matrix & classification report of all algorithms

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.
   Answer:
   Problem Type: Classification Problem
   Step by Step logic:
   - Load the customer_insurance dataset
   - Split input (driving_history, demographics) and output (insurance claim - yes/no)
   - Split train and test dataset
   - Model creation using classification algorithms.

Must try algorithm: Binary classification (SVM, logistics regression, etc.)

- Test the model
- Model evaluation using confusion matrix & classification report
- Finalize the best model for the customer insurance claim prediction based on the performance analyses from confusion matrix & classification report of all algorithms

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.
   Answer:
   Problem Type: Clustering Problem
   Step by Step logic:
   - Load the movie_recommendation dataset
   - Select the important features for X input data - (viewing preferences, watch history)
   - Data preprocessing – Standardization
   - Model creation & prediction of Y_labels [cluster group] – (using clustering algorithms: K means, agglomerative, DBSCAN, optics, birch etc.)
   - Create a new dataset with the cluster group & save it as 'cluster.csv' file
   - Visualize the clusters using seaborn.lmplot – (for visual inspection of clusters)
   - Cluster model evaluation using cluster evaluation metrics
   - Finalize the best cluster model for the problem statement – (by analyzing the cluster outputs, evaluation metric, domain knowledge and visual cluster plot)

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.
   Answer:
   Problem Type: Regression Problem
   Step by Step logic:
   - Load the surgery_patients dataset
   - Convert the categorical data (life style habits-physical activity column) into numerical data using pd.get_dummies
   - Split input (age ; medical_history – diabetes, BMI, hyper tension; life style habits-physical activity columns) and output (patients_recovery_time)
   - Split train and test dataset
   - Model creation using regression algorithms (linear regression, SVR, decision tree, random forest, boosting etc…)
   - Test the model prediction
   - Evaluate the model using r2_score

   Finalize the best model (by comparing all algorithm performance for the dataset)

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.

Answer:

Problem Type: Regression Problem

Step by Step logic:

- Load the student dataset
- Convert the categorical data (past academic performance – Grades column) into numerical data using pd.get_dummies
- Split input (study hours, attendance, past academic performance columns) and output (student's final exam score)
- Split train and test dataset
- Model creation using regression algorithms (linear regression, SVR, decision tree, random forest, boosting etc…)
- Test the model prediction
- Evaluate the model using r2_score

Finalize the best model (by comparing all algorithm performance for the dataset)