Capstone Project Scenario: Customer Behavior Intelligence System

Background:

You are part of the data science team at a digital services company. The company manages a platform that offers e-government services and tracks various behavioral features of users such as interaction frequency, session length, number of services used, and support requests.

Your management has no predefined segments or user types. They want to leverage data to:

1. Identify meaningful user segments automatically.
2. Classify new users into the correct segment as early as possible.
3. Predict how much a user is expected to spend on the platform.

As a data scientist, your task is to design an end-to-end AI system that includes all of these components.

Your Mission (Project Objectives):

You are required to implement a full data science pipeline that includes:

1. Unsupervised Clustering - Identify hidden user groups based on behavior.
2. Classification Model - Assign new users to the discovered clusters.
3. Regression Model - Predict expected user spending.
4. Visualization & Interpretation - Provide business-level explanations.
5. (Optional) Deployment Readiness - Simulate an API or dashboard presentation.

Dataset:

- Provided as: customer_behavior_unsupervised.csv
- Rows = Users
- Columns = 5 anonymized behavioral features (feature_0 to feature_4)

- No labels (unsupervised data)

Tasks Breakdown:

Step 1: Clustering (Unsupervised Learning)

- Load and explore the dataset.
- Apply K-Means or another clustering algorithm.
- Decide the optimal number of clusters (Elbow method / Silhouette Score).
- Assign cluster labels to the data.
- Describe the characteristics of each cluster in plain English.

Step 2: Classification (Supervised Learning)

- Use the newly created cluster labels as targets.
- Train a classification model (e.g., Random Forest, SVM).
- Evaluate the model's performance on test data.
- Create a function that predicts the cluster for any new user.

Step 3: Prediction (Regression)

- Simulate a numeric column such as "user_spending" using synthetic logic.
- Train a regression model (e.g., Linear Regression or XGBoost) to predict spending.
- Evaluate using R^2 Score and RMSE.

Step 4: Visualization & Reporting

- Visualize the clusters in 2D (PCA or t-SNE).
- Plot feature importances from the classification and regression models.
- Summarize your findings in a readable way for non-technical stakeholders.

(Optional) Step 5: Bonus Challenge

- Deploy your model using Streamlit, FastAPI, or Jupyter widgets.
- Build a mini-dashboard to show:
  - Input fields for user features
  - Predicted cluster and spending
  - Cluster descriptions and recommendations

Expected Deliverables:

- Python Notebook (.ipynb) or script with:
  - Clean, well-commented code
  - Visuals and plots
  - Business-friendly summary of findings
- Optionally: Streamlit app or PowerPoint report

Learning Outcomes:

- Apply real-world clustering and classification techniques.
- Simulate data in practical AI pipelines.
- Understand end-to-end workflow from data to insights to application.
- Translate technical results into business decisions.