**Question 1 : Adult Dataset**

Load Dataset1 and do any 6 of the following:

1. Calculate the mean, median, and standard deviation of the "age" and "hours-per-week" columns. What do these values suggest about the distribution of ages and working hours in the dataset?
2. Find the distribution of individuals across different "workclass." Which workclass category is the most common?
3. Analyze the frequency distribution of "education" levels. What is the most common education level, and how does it correlate with "salary" (<=50K or >50K)?
4. How many unique "native-country" values are there in the dataset? Which country has the most individuals represented?
5. Create a scatter plot of "capital-gain" vs. "capital-loss," color-coded by "salary." What does the plot reveal about the relationship between capital gains, capital losses, and income?
6. Investigate the relationship between "education-num" and "salary." Do individuals with higher "education-num" values tend to earn more (>50K)?
7. Analyze the relationship between "marital-status" and "salary." Which marital status category has the highest proportion of individuals earning >50K?
8. Analyze the "hours-per-week" distribution for different "occupation" categories. Which occupations tend to have longer working hours?
9. Investigate if there is a significant difference in "hours-per-week" between males and females.
10. Create a bar chart showing the number of individuals in each "workclass" category. Which workclass has the highest representation?

**Question 2 : Telecom Dataset**

**Load Dataset2 and do the following**

Use the given dataset to build a machine learning model for predicting customer churn ("Churn"). The goal is to predict whether a customer is likely to churn (TRUE) or not (FALSE).

1. **Data Preprocessing:**

Outline the steps you would take to preprocess the data before training the model. Consider handling categorical features (e.g., "State," "International plan," "Voice mail plan"), normalizing numerical features, and dealing with any missing values.

1. **Model Selection:**

Choose at least three machine learning algorithms (e.g., Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine) to train on this classification task. Justify your choice of models.

1. **Feature Engineering:**

Consider creating new features based on the existing ones (e.g., total minutes or total charge across all time periods, or call duration ratios). Explain how the newly engineered features could help improve the model's performance.

1. **Model Evaluation:**

Evaluate the models using appropriate metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score. Which model performs the best, and why? Provide a confusion matrix to support your evaluation.

1. **Feature Importance Analysis:**

Use feature importance techniques (e.g., coefficients from Logistic Regression, feature importance from tree-based models) to identify the most significant features for predicting churn. How do features such as "Total day minutes," "Customer service calls," or "International plan" impact the predictions?

1. **Hyperparameter Tuning:**

Perform hyperparameter tuning on the best-performing model using techniques like grid search or random search. What hyperparameters did you tune, and how did it affect the model's performance?

1. **Deployment Considerations:**

If you were to deploy this model for real-world use, what additional steps would you take to monitor its performance and maintain its accuracy over time?

**Question 3 : Telecom Dataset (Optional Question)**

**Build a Flask application for the best performance model found above.**