Bank Marketing Classification Task Assignment 🏦💰

Objective

The goal of this assignment is to build a classification model to predict whether a client will subscribe to a term deposit (`yes/no`) based on the data from a Portuguese bank's marketing campaigns.

Dataset Overview

The dataset contains information about:

1. **Client Data**: Personal details like age, job, marital status, and balance.
2. **Last Contact Details**: Information about the last contact of the current marketing campaign.
3. **Campaign Data**: Details about the number and outcomes of previous and current marketing contacts.
4. **Target Variable**: Whether the client subscribed to a term deposit (`y: yes/no`).

You can download the dataset from Kaggle.

Steps to Complete the Assignment

Step 1: Problem Statement and Setup

1. Clearly define the problem: Predict whether a client will subscribe to a term deposit (`y`).
2. Load the dataset into a Pandas DataFrame.

Step 2: Exploratory Data Analysis (EDA)

1. Univariate Analysis

- **Numeric Features**:
  - Visualize distributions using histograms for features like `age`, `balance`, and `duration`.
  - Calculate and interpret summary statistics (mean, median, min, max, standard deviation).
- **Categorical Features**:
  - Visualize frequencies using bar plots for features like `job`, `marital`, `education`, `contact`, `poutcome`, and `y`.

2. Bivariate Analysis

- **Analyze the target variable (`y`) in relation to other features**:
  - Compare `balance`, `duration`, and `campaign` across `y = yes` and `y = no` using box plots.
  - Create bar charts to observe the impact of `education`, `job`, and `housing` on the subscription outcome.

3. Multivariate Analysis

- **Interaction Analysis**:
  - Use pair plots to study interactions between numeric features like `balance`, `age`, `duration`, and `campaign`.
  - Heatmap to visualize correlations among numeric features.

4. Missing Values and Outliers

- **Identify Missing/Unknown Values**:
  - Count and analyze “unknown” values in categorical features like `job`, `education`, and `contact`.
  - Decide whether to impute or drop these rows.
- **Outlier Detection**:
  - Use box plots to identify outliers in numeric features like `balance` and `duration`.

Step 3: Data Preprocessing

1. **Handling Missing or "Unknown" Values**:
    - Impute missing values or create a new category for “unknown” in categorical features.
2. **Encoding Categorical Features**:
    - Apply one-hot encoding for categorical features like `job`, `marital`, `education`, `contact`, and `poutcome`.
3. **Feature Scaling**:
    - Normalize numeric features like `balance`, `duration`, and `campaign` for better model performance.
4. **Derived Features**:
    - Create new features such as:
        - **Contact Rate**: Number of contacts divided by `duration`.
        - **New Client Indicator**: A binary feature where `pdays = -1` means the client is new.

Deliverables

1. **EDA Report**: Include all visualizations and insights derived from univariate, bivariate, and multivariate analysis.
2. **Preprocessed Dataset**: Provide a clean and transformed dataset ready for modeling.
3. **Code**: Submit the Python code used for EDA and preprocessing in a Jupyter Notebook or Python script.