**Problem Statement: Predicting Term Deposit Subscription in Bank Marketing**

**Dataset Overview:**

You are provided with a dataset containing information about clients' interactions with a bank's marketing campaign. The target variable is binary, indicating whether a client subscribed to a term deposit or not (1 for yes, 0 for no). The dataset includes various features such as client demographics, contact details, and historical campaign information. Main objective is to classify clients into those who subscribe to a term deposit and those who do not.

Dataset: in bank\_full.csv format (uploaded as separate file)

Metadata file

The metadata file contains information about the features in the dataset, including:

1. Age: Age of the client.
2. Job: Type of job.
3. Marital: Marital status.
4. Education: Level of education.
5. Default: Has credit in default (yes, no, unknown).
6. Balance: Average yearly balance
7. Housing: Has housing loan (yes, no, unknown).
8. Loan: Has personal loan (yes, no, unknown).
9. Contact: Contact communication type.
10. Month: Last contact month of the year.
11. Day of Week: Last contact day of the week.
12. Duration: Last contact duration, in seconds.
13. Campaign: Number of contacts performed during this campaign.
14. Pdays: Number of days that passed by after the client was last contacted from a previous campaign.
15. Previous: Number of contacts performed before this campaign.
16. Poutcome: Outcome of the previous marketing campaign.
17. Target: Binary variable indicating term deposit subscription (1 = yes; 0 = no).
18. **Import Libraries/Dataset** 
    1. Download the dataset.
    2. Import the required libraries.
19. **Data Visualization and Exploration [1M]** 
    1. Print 2 rows for sanity check to identify all the features present in the dataset and if the target matches with them.
    2. Provide appropriate data visualizations to get an insight about the dataset.
    3. Do the correlational analysis on the dataset. Provide a visualization for the same. Will this correlational analysis have effect on feature selection that you will perform in the next step? Justify your answer. **Answer without justification will not be awarded marks.**
20. **Data Pre-processing and cleaning [2M]** 
    1. Do the appropriate pre-processing of the data like identifying NULL or Missing Values if any, handling of outliers if present in the dataset, skewed data etc. Mention the pre-processing steps performed in the markdown cell.
    2. Apply appropriate feature engineering techniques. Apply the feature transformation techniques like Standardization, Normalization, etc. You are free to apply the appropriate transformations depending upon the structure and the complexity of your dataset. Provide proper justification. **Techniques used without justification will not be awarded marks**. Explore a few techniques for identifying feature importance for your feature engineering task.
21. **Model Building [5M]**
    1. Split the dataset into training and test sets. **Answer without justification will not be awarded marks.** [1M]
       1. Train = 80 % Test = 20%
       2. Also, try to split the dataset with different ratios of your choice.
    2. Build model using Logistic model and decision tree [4 M]
       1. Tune hyperparameters (e.g., number of trees, maximum depth) using cross-validation. Justify your answer.
22. **Performance Evaluation [2M]**
    1. Compare the performance of the Logistic Regression and Decision Tree models using appropriate evaluation metrics.
    2. Provide insights into which model performs better and why. **Answer without justification will not be awarded marks.**

***For clarifications, contact Abinaya Marimuthu <abinaya.m@wilp.bits-pilani.ac.in>***