**Project 4 – Group 1 Project Proposal**

**Project Overview:**

The primary objective of this project is to analyze and predict customer responses to a bank’s marketing campaigns.

This project involves developing a **classification machine learning model** to predict whether a customer will subscribe to a term deposit based on their demographic and campaign data.

The insights from this analysis can help the bank optimize its marketing strategies, improve customer targeting, and enhance campaign effectiveness.

**Dataset: Bank Marketing**

Dataset is obtained from the UCI Machine Learning Repository. The dataset consists of information about marketing campaigns of a Portuguese banking institution. The data is related to direct marketing campaigns (phone calls) and contains 41,188 records and 21 attributes.

<https://archive.ics.uci.edu/dataset/222/bank+marketing>

**Target Variable:** `y` binary variable indicating whether the client subscribed to a term deposit: "yes" or "no"

**Features:** Attributes include demographic and campaign information.

**Research Question:**

1. Can we predict whether a customer will respond to marketing campaign based on the feature variables in the dataset? – **Classification Model**

**Tasks:**

1. Read Csv , Load to Dataframe , Clean and Load into Spark SQL - **Eriel**
2. Read from Spark SQL and Run few queries , visualizations - **Aliyu**
3. A Python script initializes, trains, and evaluates a model - **David**
4. Model Optimizations – **Indu**
5. ReadMe – **Everyone**
6. Slides and Presentations - **Everyone**

**# Project 4 Rubric**

**Data Model Implementation (25 points)**

* A Python script initializes, trains, and evaluates a model (10 points)
* The data is cleaned, normalized, and standardized prior to modeling (5 points)
* The model utilizes data retrieved from SQL or Spark (5 points)
* The model demonstrates meaningful predictive power at least 75% classification accuracy or 0.80 R-squared. (5 points)

**Data Model Optimization (25 points) - David**

* The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself (15 points)
* Overall model performance is printed or displayed at the end of the script (10 points)

**GitHub Documentation (25 points)**

* GitHub repository is free of unnecessary files and folders and has an appropriate .gitignore in use (10 points)
* The README is customized as a polished presentation of the content of the project (15 points)

**Group Presentation (25 points)**

* All group members speak during the presentation. (5 points)
* Content, transitions, and conclusions flow smoothly within any time restrictions. (5 points)
* The content is relevant to the project. (10 points)
* The presentation maintains audience interest. (5 points)