INF80051 Assignment 3 (Group Project)

Word count: 4000 +/- 10% words; **Due Date:** Check Canvas;
Weight: 35% (35 marks); **Submission:** word file

1 Overview

In the competitive banking industry, identifying potential customers for term deposits is crucial for revenue growth. This assignment focuses on applying AI techniques to predict whether bank clients will subscribe to a term deposit based on direct marketing campaign data. You will work with real data from a Portuguese banking institution, analyzing the effectiveness of phone-based marketing campaigns and developing predictive models to improve targeting strategies.

The business challenge is clear: banks spend significant resources on marketing campaigns with traditionally low conversion rates (around 11%). By accurately predicting which clients are likely to subscribe, banks can optimize their marketing efforts, reduce costs, and improve customer satisfaction by avoiding unnecessary contacts.

2 Your Task

You will analyze the Bank Marketing dataset to predict term deposit subscriptions. Your specific tasks include:

1. Clean and preprocess the data, addressing missing values, outliers, and the significant class imbalance (only 11.3% positive responses)
2. Develop multiple predictive models using techniques discussed in class (e.g., logistic regression, decision trees, neural networks, ensemble methods)
3. Evaluate models using appropriate metrics for imbalanced datasets (not just accuracy)
4. Provide actionable business recommendations for improving marketing campaign effectiveness

You must justify all data cleaning decisions and model choices. If using advanced techniques not covered in class, explain your rationale. Submit a comprehensive report with R code in the appendix.

3 Dataset Information

Dataset: Bank Marketing Dataset (UCI Machine Learning Repository)
Download Link: `https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing`
Alternative: `https://archive.ics.uci.edu/dataset/222/bank+marketing`
File: bank-additional-full.csv (41,188 records, 20 input features + 1 target)

3.1 Key Features

Client Data: age, job, marital status, education, credit default, housing loan, personal loan
Campaign Data: contact type, last contact date, duration, number of contacts
Economic Context: employment variation rate, consumer price index, consumer confidence index, euribor rate, number of employees
Target Variable: y (has the client subscribed a term deposit? yes/no)

Important Note: The dataset is imbalanced with only 11.3% positive cases, requiring careful handling in your analysis.

4 Proposed Structure

While no specific structure is mandated, consider organizing your report around:

1. Introduction: Business context and problem significance
2. Data Understanding: Exploration and quality assessment
3. Data Preprocessing: Cleaning techniques and justifications
4. Modeling: Multiple approaches with business rationale
5. Evaluation: Comprehensive metrics and business impact
6. Recommendations: Actionable insights for marketing strategy