DA5401 A6: Imputation via Regression for Missing Data

Objective: This assignment challenges you to apply linear and non-linear regression to impute missing values in a dataset. The effectiveness of your imputation methods will be measured indirectly by assessing the performance of a subsequent classification task, comparing the regression-based approach against simpler imputation strategies.

---

1. Problem Statement

You are a machine learning engineer working on a credit risk assessment project. You have been provided with the **UCI Credit Card Default Clients Dataset**. This dataset has missing values in several important feature columns. The presence of missing data prevents the immediate application of many classification algorithms.

Your task is to implement three different strategies for handling the missing data and then use the resulting clean datasets to train and evaluate a classification model. This will demonstrate how the choice of imputation technique significantly impacts final model performance.

You will submit a Jupyter Notebook with your complete code, visualizations, and a plausible story that explains your findings. The notebook should be well-commented, reproducible, and easy to follow.

Dataset:

- **UCI Credit Card Default Clients Dataset (with missing values):** Kaggle - Credit Card Default Clients Dataset
    - *Note: While the original UCI dataset is relatively clean, for this assignment, you should **artificially introduce Missing At Random (MAR) values** (e.g., replace 5% of the values in the 'AGE' and 'BILL_AMT' columns with NaN) before starting Part A, to simulate a real-world scenario with a substantial missing data problem.*

---

2. Tasks

Part A: Data Preprocessing and Imputation [20 points]

1. **Load and Prepare Data [4]:** Load the dataset and, as instructed in the note above, **artificially introduce MAR missing values** (5-10% in 2-3 numerical feature columns). The target variable is 'default payment next month'.
2. **Imputation Strategy 1: Simple Imputation (Baseline) [4]:**
    - Create a clean dataset copy (Dataset A).

    - For each column with missing values, fill the missing values with the **median** of that column. Explain why the median is often preferred over the mean for imputation.
3. **Imputation Strategy 2: Regression Imputation (Linear) [6]:**
    - Create a second clean dataset copy (Dataset B).
    - For a single column (your choice) with missing values, use a **Linear Regression** model to predict the missing values based on all other non-missing features. Explain the underlying assumption of this method (Missing At Random).
4. **Imputation Strategy 3: Regression Imputation (Non-Linear) [6]:**
    - Create a third clean dataset copy (Dataset C).
    - For the same column as in Strategy 2, use a **non-linear regression model** (e.g., K-Nearest Neighbors Regression or Decision Tree Regression) to predict the missing values.

Part B: Model Training and Performance Assessment [10 points]

1. **Data Split [3]:** For each of the three imputed datasets (A, B, C), split the data into training and testing sets. Also, create a fourth dataset (Dataset D) by simply **removing all rows** that contain any missing values (Listwise Deletion). Split Dataset D into training and testing sets.
2. **Classifier Setup [2]:** Standardize the features in all four datasets (A, B, C, D) using `StandardScaler`.
3. **Model Evaluation [5]:** Train a **Logistic Regression** classifier on the training set of each of the four datasets (A, B, C, D). Evaluate the performance of each model on its respective test set using a full **Classification Report** (Accuracy, Precision, Recall, F1-score).

Part C: Comparative Analysis [20 points]

1. **Results Comparison [10]:** Create a summary table comparing the performance metrics (especially F1-score) of the four models:
    - Model A (Median Imputation)
    - Model B (Linear Regression Imputation)
    - Model C (Non-Linear Regression Imputation)
    - Model D (Listwise Deletion)
2. **Efficacy Discussion [10]:**
    - Discuss the trade-off between **Listwise Deletion (Model D)** and **Imputation (Models A, B, C)**. Why might Model D perform poorly even if the imputed models perform worse?

    - Which regression method (Linear vs. Non-Linear) performed better and why? Relate this to the assumed relationship between the imputed feature and the predictors.
    - Conclude with a recommendation on the best strategy for handling missing data in this scenario, justifying your answer by referencing both the classification performance metrics and the conceptual implications of each method.

---

3. Submission Guidelines

- The assignment is due in **a week**.
- Submit a single Jupyter Notebook with all your code, visualizations, and answers to the conceptual questions in markdown cells.
- Ensure your code is clean, readable, and reproducible.

Evaluation Criteria:

- Correct implementation of the four missing data strategies.
- Clear justification for the choice of regression models.
- Accurate training and evaluation of the Logistic Regression classifier.
- Insightful comparative analysis of the final classification performance.

Good luck!