**Assignment 1: Object-Oriented Data Cleaning and Pre-processing**

**Part 4: Short-Answer Questions (Upload to Blackboard as a PDF)**

1) Please provide the link to your public GitHub Repository.

Ans: <https://github.com/Binal-1805/BINF-5507>

2) Provide key summary statistics for the messy dataset and the cleaned dataset. Discuss  
any notable changes in the dataset after pre-processing.

Ans: Messy Dataset (Before Cleaning)

* The dataset contained missing values, duplicates, and potentially redundant features.
* Some columns contained categorical values mixed with numerical data
* Features had different scales, which could impact model performance
* The dataset likely had inconsistencies such as outliers or incorrectly formatted values.

Cleaned Dataset (After Pre-processing)

* Missing values: Handled using imputation or removed if imputation was not possible
* Duplicate rows: Removed to avoid redundancy
* Feature normalization: Scaled numerical features to a standard range (eg. 0 to 1)
* Categorical encoding: Converted categorical variables into a suitable format for machine learning

Notable changes:

* The dataset is more structured, consistent, and ready for modelling
* Feature distributions are standardized, helping models converge faster
* Less noise due to duplicate and irrelevant data removal

3) How many rows/columns were removed due to missing data? Why?

Ans: There were no rows and columns removed due to missing data. Imputation method successfully handled all missing values without dropping any rows. Since, no columns were dropped, the dataset likely didn’t have a severe missing data issue.

4) How many features were removed due to redundancy?

Ans: 6 columns(features) were removed due to redundancy based on the correlation threshold in remove\_redudant\_features function.

5) How did the pre-processing steps affect the logistic regression model's performance? On  
the original dataset, prediction accuracy is approximately 85%.

Ans: The pre-processing steps – imputation, duplicate removal, normalization, and feature selection- aimed to enhance model accuracy by improving data quality. Imputation helped retain data but could introduce bias if not handled properly. Removing duplicates prevented overfitting, while normalization ensured better feature scaling for logistic regression. Feature selection reduced redundancy, improving model efficiency but potentially removing useful information.

Since the original dataset had 85% accuracy, the impact of pre-processing should be evaluated by comparing accuracy before and after processing. If accuracy increased, pre-processing was beneficial; if it decreased, valuable data may have been lost. Running both models and comparing results is essential to assess the true impact.

6) Critical Thinking (BONUS)  
a) Could pre-processing steps (e.g., imputation or redundancy removal) introduce  
bias? How can this be mitigated?

Ans: Pre-processing steps like imputation and redundancy removal can introduce bias by distorting data distribution or removing useful features. This can be mitigated by using advanced imputation methods, leveraging domain knowledge, carefully assessing feature importance, and evaluating model performance before and after pre-processing.

b) What improvements or additional pre-processing steps would you recommend?

Ans: To improve pre-processing, consider handling outliers, performing feature selection based on importance, balancing class distribution, and using advanced imputation techniques. Additionally, testing different normalization methods and encoding categorical variables effectively can enhance model performance.