**Assignment 1: Object-Oriented Data Cleaning and Preprocessing**

**Name: Sheona Hans Jeyakumar**

**Course: BINF-5507**

Part 4: Short-Answer Questions (Upload to Blackboard)  
1) Please provide the link to your public GitHub Repository.

<https://github.com/Sheona-Hans/BINF-5507-Materials.git>

Assignment folder: cd Assignment1

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

Messy\_Data:

* Rows: 1,196 total
* Target values: Only 920 non-null → 23% (approx) missing
* Numerical columns: Many float columns (e.g., b, c, ..., z) with missing values
* Column ranges: Wide spread in values
* Example: h ranges from 0 to 603; j ranges from -5.95 to 1.37
* Possible outliers in columns like: h, j, k, z — due to unusually large/small min/max
* Data Types: Mix of binary targets, continuous features, some skewed

Clean\_Data:

* Rows: 372 remaining → 69% (approx) of rows removed (likely due to NaNs in target or features)
* Columns: 22 (some dropped during redundancy)
* All missing values handled: No nulls remain
* Normalized scale: All values transformed to a consistent 0–1 or standardized range
* Example: j now ranges from 0.13 to 0.90, instead of -5.95 to 1.37
* Outliers mitigated: No extreme values remain due to scaling and filtering
* More compact and model-ready: All features are numeric and clean.
* NOTE: Target column is converted to categorical while executing simple\_model function as it uses LogisticRegression

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

* Original rows: 1,196
* Cleaned rows: 372
* Rows removed: 1,196 - 372 = 824
* Why? - Rows with missing target values, and likely rows with missing values in important features were dropped during cleaning

4) How many features were removed due to redundancy?

* Original features: 28
* Final features: 22
* Removed features: 6
* Why? - Redundant features are typically dropped using correlation thresholding

5) How did the preprocessing steps affect the logistic regression model's performance? Be  
quantitative! On the original dataset, prediction accuracy is approximately 85%.

* Slight improvement in accuracy (from 85% to 85.33%), indicating:
* Cleaned, normalized features improved model consistency
* Removing noise (outliers, NaNs) reduced variance
* While the improvement is small, it's significant in terms of model reliability and generalizability — the model is now less likely to overfit.

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

* Yes preprocessing steps can introduce bias.
  + Imputation - Filling missing values using mean/median can distort distributions and hide true variability, especially if missingness is not random.
  + Redundancy removal - Removing correlated features can discard meaningful context or interaction effects, especially if done purely based on correlation.
  + Row removal using dropna - Removing rows with missing data can skew class balance or eliminate underrepresented populations.
* How to mitigate bias
  + Use conditional imputation - Impute missing values based on related features or groups
  + Analyze missingness pattern
  + Retain missing but important features using domain knowledge and understanding correlation.

b) What improvements or additional preprocessing steps would you recommend?

* Feature Engineering
* Dimensionality reduction
* Cross-validation
* Check class balance and handle it.