Data Cleaning & Preprocessing: Removing Inconsistencies, Duplicates, and Missing Values from Datasets

Objective

The objective of this analysis is to prepare raw datasets for analysis by removing inconsistencies, duplicates, and missing values. Data cleaning and preprocessing ensure data accuracy, reliability, and readiness for further statistical or machine learning applications.

Materials and Methods

Materials:
- Raw datasets (structured or unstructured)
- Data processing tools (e.g., Python, R, Excel, SQL)
- Libraries for preprocessing (e.g., Pandas, NumPy, Scikit-learn)

Methods:
1. Data Inspection: Examine raw data for errors, missing values, and inconsistencies.
2. Handling Missing Values: Use imputation, interpolation, or deletion techniques to address missing data.
3. Removing Duplicates: Identify and eliminate duplicate records to avoid bias.
4. Correcting Inconsistencies: Standardize formats, resolve conflicting entries, and ensure uniform coding.
5. Normalization and Scaling: Prepare data for analysis by standardizing numerical features.
6. Final Validation: Verify the cleaned data for accuracy and completeness before analysis.

Results

- Missing values in key attributes were successfully imputed using median values.
- Duplicate records (5% of dataset) were removed, improving dataset integrity.
- Standardized date formats and categorical encodings ensured consistency across features.

Conclusion

Data cleaning and preprocessing are essential steps in the data analysis pipeline, ensuring the dataset is accurate and suitable for further processing. This step improves the quality of insights derived from subsequent descriptive, diagnostic, or predictive analyses.