Notes for Data Cleansing (Part I)

Personal Note

Department of Computer Science and Engineering
Green University of Bangladesh

Md. Nazmul Hossain
ID: 223002089
Semester: 6th

Data Preprocessing for Machine Learning 😊

Table of Content

1. Title:

- Data Preprocessing Project 01 | House Selling Dataset

2. Introduction of Data Preprocessing:

Data preprocessing is a crucial step in the data science process that involves cleaning, transforming, and organizing raw data into a suitable format for analysis and modeling. Real-world data is often incomplete, inconsistent, and contains errors, which can significantly impact the accuracy and reliability of any insights or models derived from it. Data preprocessing techniques address these issues by:

1. **Data Cleaning:** This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

2. **Data Integration:** This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

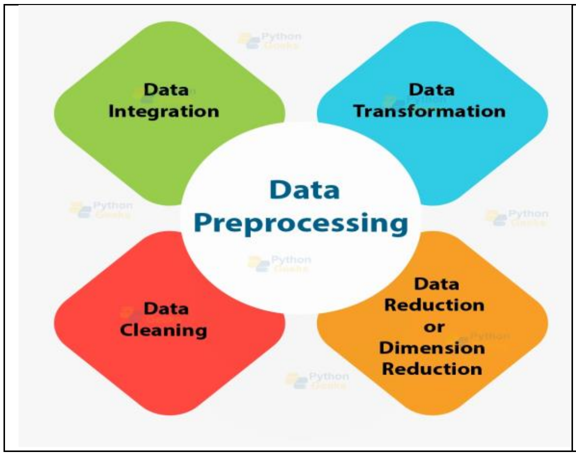
3. **Data Transformation:** This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

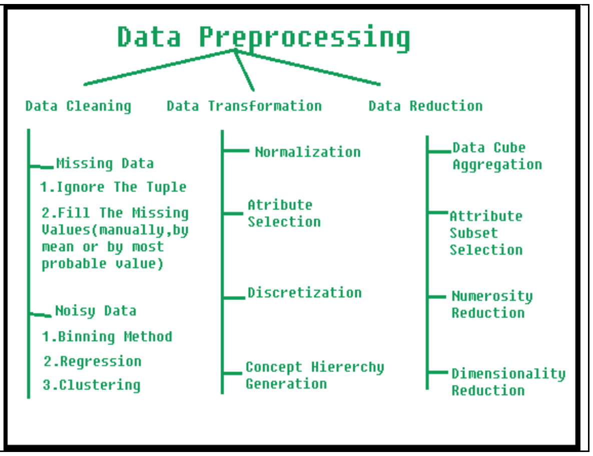
4. **Data Reduction:** This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

5. **Data Discretization:** This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

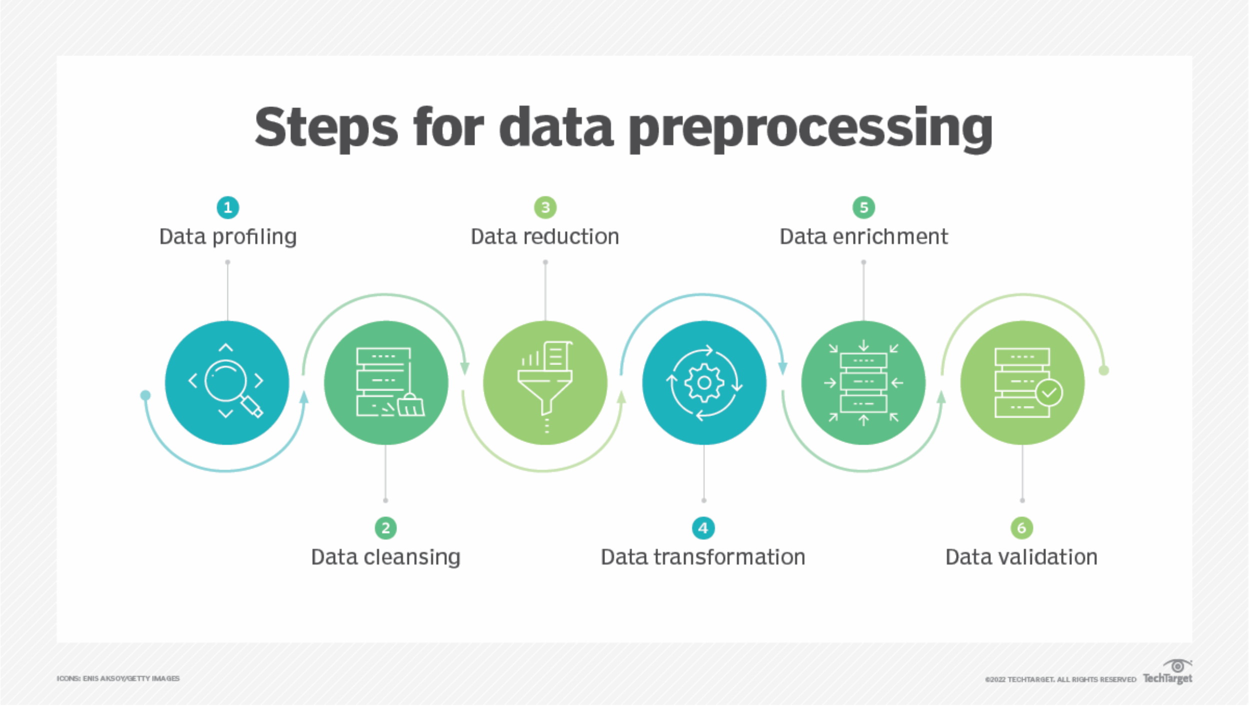
6. **Data Normalization:** This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

By preprocessing data, we can improve its quality, making it more suitable for machine learning algorithms and ensuring that the results obtained are meaningful and reliable.





3. Methodology of Data Preprocessing:



ICONS: ENIS AKSOY/GETTY IMAGES

©2022 TECHTARGET. ALL RIGHTS RESERVED TechTarget

4. Introduction of Data Cleansing:

Data cleansing, also known as data scrubbing or cleaning, is the process of identifying and correcting errors, inconsistencies, duplicates, and missing values within a dataset, aiming to improve the overall quality and reliability of the data by ensuring its accuracy and consistency before analysis or usage.

Key points about data cleansing:

- **Purpose:** To eliminate "dirty" data, including typos, incorrect formatting, missing information, and outliers, to produce a clean and usable dataset for further analysis.

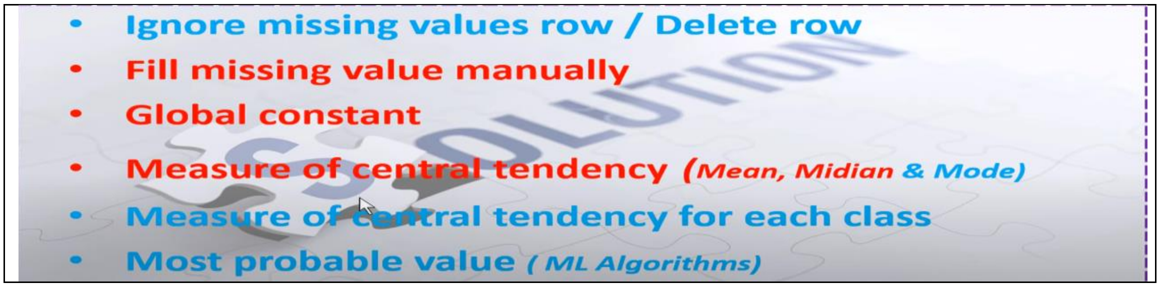
- **Common tasks:**
  - **Identifying duplicates:** Removing redundant data entries.
  - **Handling missing values:** Imputing missing data with appropriate values based on context or removing incomplete records.
  - **Standardizing formatting:** Ensuring data is presented consistently (e.g., consistent date formats, proper capitalization).
  - **Validating data:** Checking if data falls within expected ranges or adheres to predefined rules.
  - **Correcting errors:** Manually fixing incorrect data entries like typos or misspellings.
- **Importance:**
  - **Accurate analysis:** Clean data leads to reliable insights and better decision-making based on accurate results.
  - **Improved efficiency:** Reduces time spent on data preparation and troubleshooting issues.
  - **Data integrity:** Ensures data is consistent and trustworthy across different systems.

5. Process of Data Cleansing:

a. Import Datasets
b. Merge datasets into dataset
c. Identify errors
d. Exploratory Data analysis (EDA)
e. Handling missing values
f. Duplicated data removal
g. Standardize Processes
h. Data transformation
i. Validate accuracy

6. Methodology of Data Cleansing:

a) Ignore missing values row / Delete row
b) Fill missing value manually
c) Global constant
d) Measure of central tendency (Mean, Median & Mode)
e) Measure of central tendency for each class
f) Most probable value (ML Algorithms)



- Ignore missing values row / Delete row
- Fill missing value manually
- Global constant
- Measure of central tendency *(Mean, Midian & Mode)*
- Measure of central tendency for each class
- Most probable value *(ML Algorithms)*

7. Ignore Missing Value/Delete Rows/Columns:

- If null values are greater than 20% then delete **Column**
- If null values are greater than 5% then delete the **Row**
- **Steps:**
  - Check null values and with histogram
  - See percentage of each columns null value
  - Select columns that contain 17% or above null values
  - Drop by that "null column list"
  - Drop the rows
  - Then check all null values are removed or not
  - Visualization of each column after and before clean
    - Check for 1 column
    - Merge histogram
    - Check for all columns
  - Check Percentage value for object type columns

8. Fill Values Manually:

- Using Global Constant
- Measures of Central Tendency
  - Mean
  - Median
  - Mode

9. Global Constant:

- Numerical data
  - 0, -1, 0.0, -1.1, 99 any number
- String Data
  - '?', 'Undefined' etc.
- Boolean Data
  - False, True

10. Measures of Central Tendency(Mean, Median, Mode):

- **Mean**
  - Average Value →
  - On linear data
- **Median**

  - Sort data by ascending
  - Total number N
  - If odd
    - Mid value
    - Median = n/2
  - If even
    - Average of mid and mid+1 value
    - Median = ((n/2) + (n+1)/2)2
- **Mode**
  - Highest frequency
  - Basically work on categorical data
- Fill value by categorical data.

11. XXX:

12. XXX:

About Author

Author: Md. Nazmul Hossain

Profession: Student

Department: Computer Science and Engineering

Institution: Green University of Bangladesh

Skill: C/C++, Java, JavaFX, Assembly Language, Python, Pandas, Numpy, DSA, CP, GitHub

Achievement: Qualify for ICPC 2024 Dhaka Regional

Social Link:

- GitHub
- X(Twitter)