**UNIT – I: About Data**

**Introduction, Causality and Experiments - Data Pre-processing: Knowing data, Data cleaning, Data reduction, Data transformation, Data discretization.**

**Data Pre-processing**

**1. Introduction to Data Pre-processing**

Data pre-processing is a crucial step in data analysis and machine learning, as raw data often contains noise, missing values, and inconsistencies. Proper pre-processing ensures data quality and enhances model performance.

**Objectives of Data Pre-processing:**

* Improve data quality
* Enhance model accuracy
* Reduce computational cost
* Ensure consistency and reliability

**2. Knowing the Data**

Understanding the dataset before processing is essential. It includes:

* **Data Types**: Numerical (integer, float), Categorical (ordinal, nominal), Text, and Date-time.
* **Data Distribution**: Mean, median, standard deviation, skewness, and kurtosis.
* **Identifying Outliers**: Boxplots, histograms, and Z-score methods.
* **Missing Values**: Identifying missing values using exploratory data analysis (EDA).

**3. Data Cleaning**

Data cleaning involves handling missing values, removing duplicates, correcting inconsistencies, and handling outliers.

**Methods of Data Cleaning:**

* **Handling Missing Data**:
  + Removing rows/columns with missing values
  + Imputation (mean, median, mode, or predictive methods)
* **Removing Duplicates**
* **Handling Inconsistent Data**: Standardizing formats, correcting spelling errors
* **Dealing with Outliers**:
  + Using statistical methods (Z-score, IQR method)
  + Transforming outliers using log or scaling techniques

**4. Data Reduction**

Data reduction helps in minimizing data complexity without losing essential information. This is especially useful in large datasets where processing all features is computationally expensive.

**Techniques of Data Reduction:**

* **Dimensionality Reduction:**
  + Principal Component Analysis (PCA): Reduces the number of dimensions while preserving variance.
    - Example: In an e-commerce dataset with 100 features (e.g., customer demographics, purchase history), PCA can reduce it to 10 principal components capturing most of the variance, speeding up model training.
  + Linear Discriminant Analysis (LDA): Focuses on maximizing class separability, used in classification problems.
    - Example: In a facial recognition system, LDA can reduce feature space while keeping essential details that differentiate faces.
* **Feature Selection:**
  + Filter Methods: Use statistical tests (e.g., chi-square test, correlation analysis) to remove irrelevant features.
  + Wrapper Methods: Select features based on model performance (e.g., recursive feature elimination).
  + Embedded Methods: Feature selection occurs during model training (e.g., Lasso regression).
  + Example: In a medical diagnosis dataset, features like 'patient ID' may not be useful for prediction and can be removed to improve model efficiency.
* **Data Sampling:**
  + Random Sampling: Selecting a subset of data randomly.
  + Stratified Sampling: Ensures proportional representation of different groups.
  + Example: If a dataset has 1 million customer transaction records, a random sample of 10,000 records can be selected to conduct a quick exploratory analysis before training a model.
* **Data Compression:**
  + Huffman Coding: A lossless compression algorithm for reducing storage requirements.
  + Wavelet Transformation: Transforms data into a compressed representation while retaining essential information.
  + Example: In image processing, wavelet transformation is used to compress high-resolution satellite images while maintaining key geographical details for analysis.

**5. Data Transformation**

Data transformation improves model efficiency and performance by modifying data into suitable formats.

**Techniques of Data Transformation:**

* **Normalization (Scaling Data):**
  + Min-Max Scaling: Rescales values between [0,1].
    - Example: In financial applications, stock prices are scaled between 0 and 1 to ensure different stocks are comparable in predictive models.
  + Z-score Normalization: Centers data to mean = 0, standard deviation = 1.
    - Example: In medical datasets, patient blood pressure values can be standardized to ensure features with different units are comparable.
* **Encoding Categorical Data:**
  + One-Hot Encoding: Converts categorical values into binary vectors.
    - Example: In customer segmentation, categories like "Gold," "Silver," and "Bronze" membership are transformed into binary values.
  + Label Encoding: Assigns numerical values to categories.
    - Example: Converting "Male" and "Female" into 0 and 1 for a machine learning model in a healthcare dataset.
* **Log Transformation: Helps in dealing with skewed data by compressing large values.**
  + Example: In income prediction, salaries have a skewed distribution, and applying a log transformation makes the data more normally distributed.
* **Box-Cox Transformation: Used for normalizing non-normal data.**
  + Example: In weather forecasting, temperature data can be transformed using the Box-Cox method to stabilize variance and improve model accuracy.
* **Power Transformation: Adjusts data distributions to improve normality.**
  + Example: Used in biomedical applications where gene expression data often follows a skewed distribution, making it easier for statistical analysis.

**6. Data Discretization**

Data discretization is converting continuous data into discrete bins or categories, useful for classification and decision trees.

**Techniques of Data Discretization:**

* **Binning:**
  + Equal-Width Binning: Dividing the range into equal intervals.
  + Equal-Frequency Binning: Ensuring each bin has the same number of data points.
  + Example: In a weather prediction dataset, temperature values can be divided into categories such as "Low," "Medium," and "High" to make trend analysis easier.
* **Histogram-Based Discretization:**
  + Uses histogram patterns to determine appropriate bins.
  + Example: In an online retail store, customer spending amounts can be categorized into different spending levels (e.g., low, medium, high) based on histogram analysis.
* **Clustering-Based Discretization:**
  + Groups similar data points into clusters before assigning discrete values (e.g., k-means clustering).
  + Example: In social media sentiment analysis, text sentiment scores can be clustered into "Negative," "Neutral," and "Positive" categories.
* **Decision Tree-Based Discretization:**
  + Decision trees are used to define splitting points for discretization.
  + Example: In medical diagnostics, patient age can be discretized into meaningful categories (e.g., "Child," "Adult," "Senior") based on a decision tree classification model.

**7. Conclusion**

Data pre-processing is a foundational step in data science and machine learning. It ensures data quality, improves model accuracy, and enhances interpretability. Understanding different pre-processing techniques is vital for effective data analysis.