🧠 UNIT – 2 : DATA PREPROCESSING (Easy English Notes)

📍 1. Why to Preprocess Data

📘 Meaning:

- **Data Preprocessing** means **cleaning and converting raw data** into a proper format before analysis.
- Real-world data is often:
  - **Incomplete** (missing values),
  - **Noisy** (contains errors/outliers),
  - **Inconsistent** (spelling or format mistakes).

📗 Examples:

- Missing value → Occupation = " "
- Error in data → Salary = "abcxy"
- Inconsistent → “Gujarat” & “Gujrat”

⚙️ Purpose:

- Follows the rule: **Garbage In → Garbage Out (GIGO)**.
  → Poor quality data = wrong results.
- Data preprocessing improves data quality for better decision-making.

✅ **Note:** Around **90% of work** in data mining is done on data cleaning and transformation.

🧹 2. Data Cleaning

📘 Meaning:

Cleans data by fixing missing, wrong, or duplicate information.

🔹 Tasks in Data Cleaning:

1. Fill missing values
2. Identify outliers and smooth noisy data
3. Correct inconsistent data
4. Remove redundancy

🧩 1) Fill Missing Values

Methods:

1. **Ignore the tuple** – delete record with missing data (if few).
2. **Fill manually** – not practical for large data.
3. **Use global constant** – replace with "Unknown" or -∞.
4. **Use Mean/Median** –
   - For normal data → use **Mean**
   - For skewed data → use **Median**
5. **Use Mean/Median of same class** – replace within that class.
6. **Use most probable value** – predict using **regression** or **decision tree**.

🧩 2) Identify Outliers and Smooth Noisy Data

Techniques:

1. **Binning** – group data into bins, replace with bin mean/median/boundary.
2. **Regression** – fit data to a line/function to find patterns.
3. **Clustering** – group similar data and detect outliers.

🧮 Example (Binning)

Data: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
→ Divide into bins (equal depth):

- Bin 1: 4, 8, 9, 15 → Mean = 9
- Bin 2: 21, 21, 24, 25 → Mean = 23
- Bin 3: 26, 28, 29, 34 → Mean = 29

Replace values with bin mean.

🧩 3) Correct Inconsistent Data

- Solve spelling/grammar issues (e.g., “Gujrat” → “Gujarat”).
- Use:
  - Domain knowledge
  - Standard formatting
  - Reference tables (master data)
  - Duplicate detection tools

🧩 4) Resolve Redundancy

- Occurs when same data is stored multiple times.
- Example: Customer info stored with every purchase record.
- Solution:
  - Use **Normalization**
  - Use **Foreign Keys**

✅ **Result:** Data becomes accurate, consistent, and non-repetitive.

🔗 3. Data Integration

📘 Meaning:

Combining data from multiple sources into one consistent dataset.

🎯 Purpose:

- Merge related data
- Remove duplicates
- Ensure same format and meaning

🧩 Schema Integration

- Match attributes from different databases.
  Example: A.cust_id ≡ B.cust#

🧩 Entity Identification Problem

- Identify real-world entities across sources.
  Example: cust_number = customer_id

🧩 Redundancy & Correlation Analysis

- **Redundancy** → storing same info twice
  Example: annual revenue = sum of monthly revenue
- **Correlation** → check relation between attributes
  - **Positive** → both increase (Study hours ↑, Marks ↑)
  - **Negative** → one increases, other decreases (Temp ↑, Hot coffee sales ↓)
  - **None** → unrelated (Shoe size vs IQ)

🧮 Chi-Square (χ^2) Test for Nominal Data

Used to check if two **categorical attributes** are related.

Steps:

1. Define hypothesis (H_0: no relation, H_1: relation exists)
2. Make contingency table (Observed data)
3. Calculate Expected values:

$$E = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

4. Find $\chi^2 = \Sigma\, (O–E)^2 / E$

Large χ^2 → strong relation between variables.

🔄 4. Data Transformation

📘 Meaning:

Converting data into another useful format for analysis.

⚙️ Methods:

Method	**Description**
Smoothing	Removes noise
Feature Construction	Create new attributes
Aggregation	Summarize data
Normalization	Scale data to smaller range
Discretization	Convert numeric to categorical

🔹 1) Min–Max Normalization

$$v' = \frac{v - \text{Min}(A)}{\text{Max}(A) - \text{Min}(A)} \times (\text{NewMax} - \text{NewMin}) + \text{NewMin}$$

Example:
Min = 16, Max = 40
Value = 30
→ $(30 - 16)/(40 - 16) = 0.58$

🔹 2) Decimal Scaling

$$v' = \frac{v}{10^j}$$

where j = smallest integer so Max(|v'|) < 1

Example:
Max = 3 → divide by 10 → values between 0 and 1.

🔹 3) Z–Score Normalization

$$v' = \frac{v - \mu}{\sigma}$$

where μ = mean, σ = standard deviation.

Example:
μ = 54,000, σ = 16,000, v = 73,600
→ z = 1.225

🔹 4) Discretization

Convert continuous → discrete intervals.
Example:
Age 10–22 = Young, 23–70 = Mature, 71–100 = Senior.

🔹 5) Concept Hierarchy

Arrange data into levels:

Country → State → City → Street

Higher level = general, lower level = detailed.

📉 5. Data Reduction

📘 Meaning:

Reducing data size but keeping important info.

🎯 Purpose:

- Faster analysis
- Less storage
- Maintain data accuracy

⚙ Techniques:

Type	Purpose
Dimensionality Reduction	Reduce attributes
Numerosity Reduction	Reduce records
Compression	Compact data storage

◆ 1) Dimensionality Reduction

Remove irrelevant attributes.

Techniques:

- **PCA (Principal Component Analysis)**
  Converts many variables → few main ones capturing most information.
  Steps: Standardize → Covariance → Eigenvalues → Sort → Transform.
- **Attribute Subset Selection**
  Keep only useful attributes (using forward/backward/decision tree).

◆ 2) Numerosity Reduction

Replace large data with smaller representation.

Techniques:

- **Histograms** – group values into bins.
- **Clustering** – group similar data, use cluster centers.
- **Sampling** – select part of data representing full dataset.

Types of Sampling:

Type	Description
SRSWOR	Random selection, no repeats
SRSWR	Random selection, repeats allowed
Cluster Sampling	Select whole clusters randomly
Stratified Sampling	Sample from each group (strata)

◆ 3) Data Compression

Store same data in smaller space.

Type	Description	Example
Lossless	100% recoverable	ZIP, PNG
Lossy	Some info lost	MP3, JPEG

⚙️ 6. Data Discretization

📘 Meaning:

Convert continuous numeric data → discrete (category) data.
E.g., Age = 23 → “Mature”.

⚙️ Techniques:

Method	Description
Binning	Divide into intervals, replace by mean/median
Histogram Analysis	Use data distribution for bins
Clustering	Group similar values
Decision Tree	Auto-split intervals
Correlation	Use related attributes to make bins

✅ Used to make data simpler and more understandable.

🌳 7. Concept Hierarchy (Detailed)

📘 Meaning:

Organize data in **multiple levels** — from detailed to summarized.

Example:

Country → State → City → Street

Level	Example	Distinct Values
1	Country	15
2	State	365
3	City	3,567
4	Street	6,74,339

⚙️ Operations:

- **Roll-up:** Move up (City → State → Country)
- **Drill-down:** Move down (Country → State → City)

✅ Helps in summarizing, comparing, and analyzing data efficiently.

🧾 📚 UNIT – 2 Summary Chart

Topic	Key Points
Preprocessing	Prepare raw data
Cleaning	Fix errors, fill missing values
Integration	Combine multiple data sources
Transformation	Normalize, scale, convert
Reduction	Make data smaller & faster
Discretization	Convert numeric → category
Concept Hierarchy	Arrange data levels for summary

✅ **FINAL RESULT:**
After all preprocessing steps:
👉 Data becomes **clean, consistent, integrated, transformed, reduced, and categorized** — ready for **data mining** and **pattern discovery**.