Data Preprocessing – Unit 2 Viva Notes

What is Data Preprocessing

Data Preprocessing means converting raw real-world data into a clean, understandable format. Real-world data is often incomplete, noisy, and inconsistent. It follows the rule: Garbage In → Garbage Out. Clean data is essential for correct results. About 90% of data mining work is data preparation.

Data Cleaning

Used to correct problems in data.

1. Filling Missing Values

1. Ignore the row (if class label missing)
2. Fill manually
3. Use constant like 'Unknown'
4. Use mean/median (mean for normal, median for skewed data)
5. Use class-wise mean/median
6. Use most probable value (regression, Bayesian, decision trees)

2. Identify Outliers & Smooth Noisy Data

Techniques:

1. Binning – Sort data, split into bins, replace with bin mean/median/boundary
2. Regression – Fit a line or curve (linear or multiple regression)
3. Clustering – Group similar data; far points are outliers

3. Correct Inconsistent Data

Standardize formats (spelling, grammar, codes), use master tables, remove duplicates.

4. Resolve Redundancy

Duplicate data wastes space and causes errors. Use normalization and foreign keys to remove redundancy.

Data Integration

Combine data from multiple sources into one coherent dataset. Helps remove redundancy and inconsistencies.

Challenges

1. Entity Identification – Match same entity from different sources (cust_id vs customer_no)
2. Redundancy and Correlation Analysis – Detect attributes derived from others
3. Use correlation coefficient for numeric and chi-square test for categorical data

Data Transformation

1. Smoothing – remove noise
2. Attribute construction – create new attributes
3. Aggregation – summarizing data
4. Normalization – scale values (Min-Max, Decimal scaling, Z-score)
5. Discretization – convert numeric to categorical

Normalization Methods

Min-Max: (v - min)/(max - min) → scales 0-1

Decimal scaling: move decimal based on max value

Z-score: (value - mean)/SD → centers to mean 0, SD=1

Data Reduction

Make data smaller but keep same meaning (faster mining).

Techniques

1. Dimensionality Reduction – fewer attributes (PCA, Attribute subset selection)
2. Numerosity Reduction – fewer records (Histograms, Clustering, Sampling)
3. Data Compression – encode data in compact form

Sampling Types

1. SRSWOR – Simple random sample without replacement
2. SRSWR – Simple random sample with replacement
3. Cluster sample – choose clusters randomly
4. Stratified sample – sample from each subgroup

Data Discretization

Convert continuous numeric data into categories. Makes data easier to understand and suitable for algorithms that need categorical data.

1. Binning (equal width/frequency)
2. Histogram Analysis
3. Clustering-based discretization
4. Decision tree-based discretization
5. Correlation-based discretization

Concept Hierarchy

Organize data from general → specific in a tree form. Example: Country → State → City → Street. Helps drill down or roll up during data analysis.