AKTU · 3RD YEAR · SEMESTER 5 · BCDS051

Business Intelligence and Analytics

Unit 3 — BI – Data Preparation

- Introduction to Data Validation
- Data Transformation – Standardization and Feature Extraction
- Data Reduction – Sampling, Selection, PCA
- Data Discretization



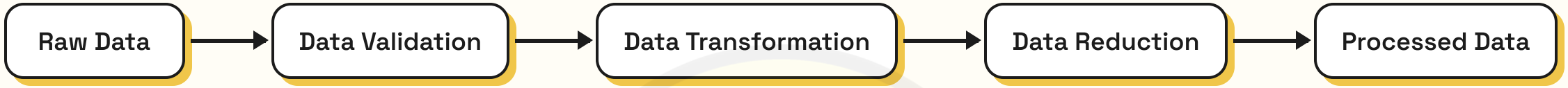
DECODED BY

Chirag Jain

Founder, Ycotes · Gold Medalist in Artificial Intelligence, IIT Madras · taught 10,000+ students

theycotes.com

Data Preparation Pipeline in BI



Overview of the Data Preparation workflow in Business Intelligence ETL processes

Introduction to Data Validation

DATA VALIDATION

Data Validation is the process of checking the accuracy, completeness, and quality of incoming raw data before it enters downstream processing, transformation, or analytical modeling.

- Ensures clean and reliable data enters the Business Intelligence pipeline.
- Prevents system failures caused by malformed inputs or missing values.
- Enforces domain constraints, schema definitions, and business logic rules.

Data Validation Techniques & Rule Checks

Common validation procedures applied during data ingestion:

- DataType Validation: Verifies that attribute values conform to expected data types (e.g., Integer, Date, Boolean).
- Range & Boundary Checking: Ensures numerical inputs fall within acceptable bounds (e.g., Age between 0 and 120).
- Cross-Field Consistency: Validates logical relations across multiple attributes (e.g., Delivery Date >= Order Date).
- Uniqueness Verification: Confirms primary identifiers (e.g., Customer ID) contain no duplicate values.
- Completeness Checking: Identifies illegal nulls in mandatory data fields.

Handling Missing and Invalid Data

Technique	Mechanism	Advantages	Disadvantages
Tuple Deletion	Remove rows containing missing attributes	Simple, clean subset	Loss of data, introduces bias
Mean/Median Imputation	Replace missing value with attribute mean/median	Preserves sample size	Distorts variance and distribution
Predictive Imputation	Use Regression or Decision Trees to predict missing values	High accuracy, keeps structure	Computationally expensive
Global Constant	Replace missing values with a fixed symbol ('Unknown')	Easy implementation	May create false categories

Data Transformation Overview

Data Transformation converts raw data into appropriate formats for analytical mining.

- Smoothing: Removes noise using binning, regression, or clustering techniques.
- Aggregation: Summarizes data values across specific dimensions (e.g., daily sales to monthly totals).
- Generalization: Replaces low-level raw attributes with higher-level concepts (e.g., numeric age to categorical age-groups).
- Normalization: Scales numeric attribute values to fall within a specified range.
- Attribute Construction: Creates new explanatory variables from existing attributes to aid predictive modeling.

Data Standardization & Normalization Formulas

Mathematical scaling techniques to achieve comparable feature magnitudes:

- Min-Max Normalization: Scales data to a target range $[new_min_A, new_max_A]$.
- Formula: $v' = \frac{v - min_A}{max_A - min_A}(new_max_A - new_min_A) + new_min_A$
- Z-Score Standardization: Transforms data to mean $\mu = 0$ and standard deviation $\sigma = 1$.
- Formula: $v' = \frac{v - \mu_A}{\sigma_A}$
- Decimal Scaling: Normalizes by moving the decimal point of attribute values.
- Formula: $v' = \frac{v}{10^j}$ where j is the smallest integer such that $Max(|v'|) < 1$.

Worked Example: Min-Max and Z-Score Normalization

An attribute Income has values ranging from $min = 12000$ to $max = 98000$, with mean $\mu = 54000$ and standard deviation $\sigma = 16000$. Normalize $v = 73000$ using Min-Max to $[0, 1]$ and Z-Score Standardization.

1. Step 1 (Min-Max): $v' = \frac{73000 - 12000}{98000 - 12000} = \frac{61000}{86000} \approx 0.7093$
2. Step 2 (Z-Score): $v' = \frac{73000 - 54000}{16000} = \frac{19000}{16000} = 1.1875$

Normalized values: Min-Max [0, 1] = 0.7093, Z-Score Standardized = 1.1875.

Feature Extraction

FEATURE EXTRACTION

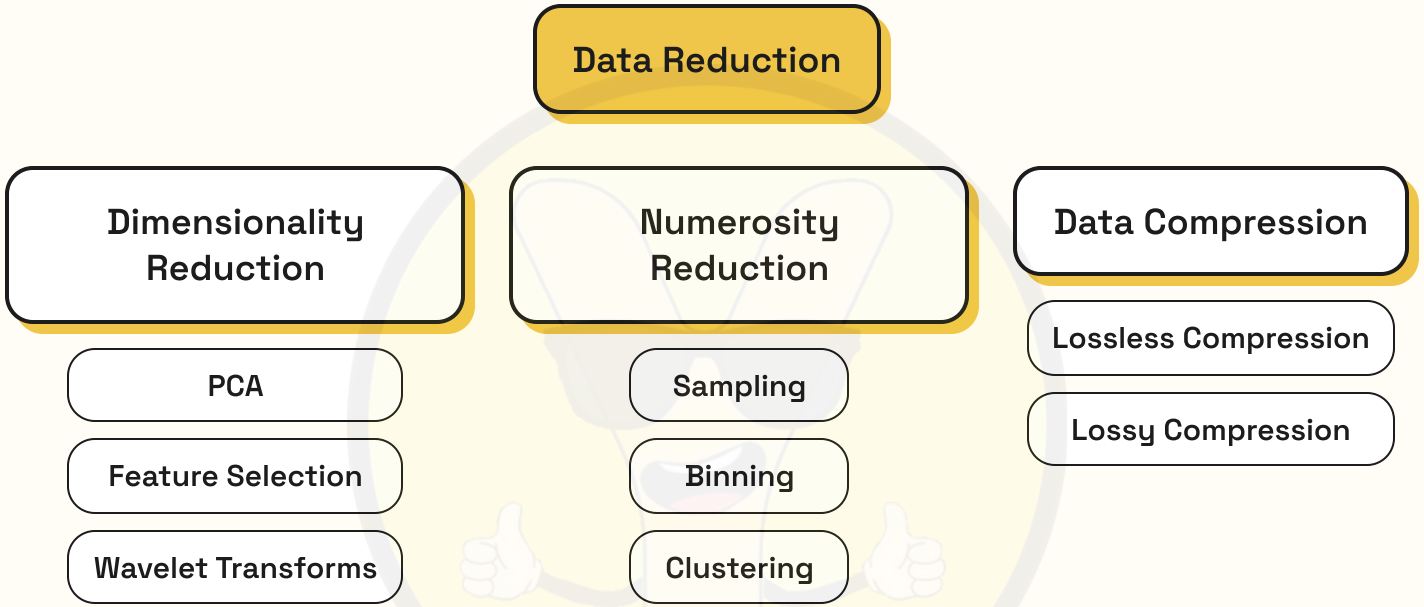
Feature Extraction is the process of mapping high-dimensional raw feature spaces into a new, reduced-dimensional feature space through mathematical combinations of original variables.

- Combines existing features into a smaller set of synthetic features.
- Reduces computational complexity and mitigates the curse of dimensionality.
- Constructs features that capture maximal variance or pattern information.

Feature Extraction Techniques

- Principal Component Analysis (PCA): Unsupervised linear transformation finding orthogonal axes of maximum variance.
- Linear Discriminant Analysis (LDA): Supervised technique projecting data to maximize class separability.
- Independent Component Analysis (ICA): Decomposes multivariate signals into additive, statistically independent components.
- Autoencoders: Neural network architectures designed to compress data into low-dimensional latent representations.
- Kernel PCA: Non-linear extension of PCA using kernel functions to map data into higher-dimensional spaces.

Data Reduction Taxonomy



Classification of Data Reduction Techniques in Data Mining

Data Reduction: Sampling Methods

Sampling reduces dataset volume by selecting a representative subset of original records.

- Simple Random Sampling Without Replacement (SRSWOR): Each record has an equal probability of selection; selected items are not returned.
- Simple Random Sampling With Replacement (SRSWR): Selected items are returned to the pool, allowing potential duplicate selections.
- Stratified Sampling: Data is partitioned into disjoint strata (e.g., by age category), and random samples are drawn from each stratum.
- Cluster Sampling: Data is divided into clusters; entire clusters are randomly chosen for analysis.
- Systematic Sampling: Selects every k-th record from an ordered dataset sequence.

Feature Selection Techniques

Feature Selection chooses a subset of relevant features without altering their original representations.

- Filter Methods: Evaluate features independently of models using statistical metrics (e.g., Chi-Square, Information Gain, Correlation Coefficient).
- Wrapper Methods: Use a predictive model to evaluate feature subsets based on performance (e.g., Forward Selection, Backward Elimination).
- Embedded Methods: Perform feature selection directly during model training (e.g., LASSO L1 Regularization, Decision Trees).
- Heuristic Search: Applies greedy strategies to avoid exhaustive evaluation of all 2^d feature subsets.

Feature Selection vs. Feature Extraction

Attribute	Feature Selection	Feature Extraction
Definition	Selects a subset of existing attributes	Constructs new features from original data
Data Representation	Original feature meanings preserved	Original features transformed to new space
Interpretability	High (retains original variables)	Low (synthetic components)
Techniques	Filter, Wrapper, Embedded	PCA, LDA, ICA, Autoencoders
Dimensionality Goal	Direct column elimination	Information compression into fewer dimensions

Principal Component Analysis (PCA)

PRINCIPAL COMPONENT ANALYSIS

PCA is an unsupervised linear dimensionality reduction algorithm that projects high-dimensional data onto a lower-dimensional subspace of orthogonal components called Principal Components, ordered by explained variance.

- First Principal Component (PC1): Direction of maximum variance in the data dataset.
- Second Principal Component (PC2): Direction of maximum remaining variance, orthogonal to PC1.
- Preserves global linear data structures while dropping low-variance components.

PCA Algorithm - Step-by-Step

Algorithmic workflow to compute Principal Components:

- Step 1: Standardize the original d-dimensional dataset to mean 0 and variance 1.
- Step 2: Compute the $d \times d$ Covariance Matrix Σ of the standardized features.
- Step 3: Calculate Eigenvalues (λ_i) and Eigenvectors (v_i) of matrix Σ.
- Step 4: Sort Eigenvectors in descending order by their corresponding Eigenvalues.
- Step 5: Select top k Eigenvectors to construct projection matrix W of size $d \times k$.
- Step 6: Transform input matrix X using $X_{new} = X \cdot W$.

Mathematical Foundations of PCA

- Covariance Matrix Formula: $Cov(X_i, X_j) = \frac{1}{n-1} \sum_{k=1}^{n} (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$
- Eigen-decomposition Equation: $\Sigma v = \lambda v$, where v is an Eigenvector and λ is an Eigenvalue.
- Explained Variance Ratio (EVR): $EVR_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}$
- Cumulative Explained Variance: Choose k principal components such that $\sum_{i=1}^{k} EVR_i \geq \text{Threshold}$ (e.g., 90%).
- Reconstruction Loss: Information lost by discarding the bottom $d - k$ components.

Data Discretization

DATA DISCRETIZATION

Data Discretization is the process of converting continuous numeric attributes into discrete intervals or categorical concepts, significantly reducing the number of distinct values for an attribute.

- Replaces numeric values with interval labels (e.g., age [0-18] mapped to 'Child').
- Necessary for data mining algorithms requiring categorical inputs (e.g., Association Rules, Decision Trees).
- Improves readability and reduces computational complexity.

Data Discretization Methods Comparison

Method	Type	Mechanism	Pros & Cons
Equal-Width Binning	Unsupervised	Divides range into N intervals of equal size	Simple; sensitive to outliers
Equal-Frequency Binning	Unsupervised	Divides data so each of N bins has equal count	Handles skewed data; distorts natural gaps
Entropy-Based Discretization	Supervised	Uses class entropy to find optimal split points	Maximizes predictive power; needs labels
Cluster-Based Discretization	Unsupervised	Applies K-Means to form natural numeric clusters	Captures data distribution; iterative process

Worked Example: Equal-Width and Equal-Frequency Binning

Given continuous values: $A = [4, 8, 15, 21, 21, 24, 25, 28, 34]$. Perform (a) Equal-Frequency Binning into 3 bins, and (b) Equal-Width Binning into 3 bins.

1. (a) Equal-Frequency (3 elements per bin): Bin 1 = [4, 8, 15], Bin 2 = [21, 21, 24], Bin 3 = [25, 28, 34].
2. (b) Equal-Width: Range = $34 - 4 = 30$. Width = $\frac{30}{3} = 10$.
3. Interval 1: $[4, 14)$ → Bin 1 = [4, 8]
4. Interval 2: $[14, 24)$ → Bin 2 = [15, 21, 21]
5. Interval 3: $[24, 34]$ → Bin 3 = [24, 25, 28, 34]

Equal-Frequency produces bins with equal item counts (3 each). Equal-Width produces bins of uniform interval length (10 each).

Exam Corner - Likely AKTU Questions

Q1 1. Define Data Validation. Explain various techniques used to handle missing and noisy data during ETL.

Q2 2. Differentiate between Min-Max Normalization and Z-Score Standardization with mathematical formulas and a numerical example.

Q3 3. What is Data Reduction? Compare Dimensionality Reduction and Numerosity Reduction with examples.

Q4 4. Differentiate between Feature Selection and Feature Extraction. Detail Filter vs Wrapper feature selection methods.

Q5 5. Explain the complete step-by-step algorithm of Principal Component Analysis (PCA). How are Eigenvalues utilized?

Q6 6. Explain Data Discretization. Compare Equal-Width and Equal-Frequency binning with a suitable numerical dataset.

Q7 7. Write short notes on: (a) Stratified Sampling, (b) Entropy-Based Discretization, (c) Attribute Construction.

Unit 3 done. You are exam-ready.

Revise the Exam corner once more, then open Unit 4 on theycotes.com. Every AKTU subject, unit by unit, free. Notes by Chirag Jain, Ycotes.

