SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name – 23ADT202 Fundamental of Data science and Analytics

II Year / IV Semester

Unit 1 – Introduction to Data science
Cleasing,integrating,and transforming data

Introduction to Data Preparation

- Raw data is often messy and unstructured.
- Accurate analysis requires thorough cleansing, integration, and transformation.
- These steps ensure that the data is usable and meaningful for exploration and modeling.
- We will explore each process: Cleansing, Integration, and Transformation.

Data Cleansing Overview

- Data cleansing detects and corrects inaccuracies.
- Key tasks: handling missing values, removing duplicates, and correcting data types.
- Clean data improves the quality and accuracy of analysis.
- Let's explore the methods in detail.

Handling Missing Values

- Missing data is common and must be handled.
- Techniques include:
  - Imputation: Replacing with mean, median, or mode.
  - Forward/Backward Fill: Using neighboring data.
  - Dropping: Removing rows/columns with excessive missing data.
- Example: Filling missing "Age" values with the mean and removing rows with missing "Salary."

Removing Duplicates

- Duplicates can distort analysis by overrepresenting records.
- It is essential to identify and remove duplicate rows in datasets.
- Example: Removing duplicate rows in a dataset with "ID" and "Name."
- Result: A cleaner, more accurate dataset for analysis.

Correcting Data Types

- Mismatched data types can lead to errors during analysis.
- Correcting data types ensures consistency.
- Example: Converting the "Age" column to numeric values.
- Potential errors may occur if the data isn't cleaned properly.

Data Integration Overview

- Data integration combines multiple datasets into one cohesive dataset.
- Ensures consistency and resolves conflicts between data sources.
- Common tasks include merging datasets, resolving conflicts, and deduplication.
- This step prepares data for holistic analysis.

Merging Datasets

Datasets are often stored separately and need to be combined.
Use common identifiers (keys) to merge data.
Example: Merging two datasets on the "ID" column to combine "Name" and "Salary" data.
Result: A unified dataset ready for further analysis.

Data Transformation Overview

- Data transformation prepares data for analysis by converting it into a suitable format.
- Techniques include normalization, scaling, encoding, and feature engineering.
- Transformed data improves model performance and accuracy.
- Let's explore some specific transformation techniques.

Transformation Techniques

Normalization: Rescaling data to a range, e.g., 0 to 1. Example: Normalizing "Salary.“

One-Hot Encoding: Converts categorical data into binary vectors. Example: Encoding "Gender.“

Feature Engineering: Creating new variables to enhance model performance. Example: Deriving BMI from weight and height.

These techniques enhance the usability of the data for analysis and modeling.

THANK YOU