Unit II

9. Data Transformation:

Data transformation in data analytics refers to the process of altering the original data to make it more suitable for analysis or modeling. Data transformation involves applying various mathematical or statistical techniques to modify the data, address specific issues, or improve the performance of analytical methods. It is a crucial step in data preprocessing that helps ensure the data's quality, normality, and compatibility with the chosen analytical techniques. Some common data transformation techniques include:

1. **Scaling and Normalization:**
    - Scaling techniques, such as Min-Max scaling or Z-score standardization, are used to bring numerical data to a similar scale. Normalization ensures that all variables have the same range, preventing variables with larger magnitudes from dominating the analysis.
2. **Log Transformation:**
    - Logarithmic transformation is applied to data with a skewed distribution, making it more symmetric and stabilizing the variance. It is commonly used when the data's variability increases with its magnitude.
3. **Power Transformation:**
    - Power transformation, such as the Box-Cox transformation, adjusts the data's distribution by applying a power function. It helps make the data more closely resemble a normal distribution, making it suitable for statistical modeling techniques that assume normality.
4. **Binning and Discretization:**
    - Binning involves dividing numerical data into discrete intervals or bins. It simplifies the data and can be helpful for analyzing patterns within specific ranges or when dealing with continuous variables in classification tasks.
5. **Handling Missing Data:**
    - Data transformation techniques are used to handle missing data points. Methods like imputation replace missing values with estimates based on other data points or statistical measures.
6. **Encoding Categorical Variables:**
    - Categorical variables are often transformed into numerical values through techniques like one-hot encoding or label encoding, enabling their use in algorithms that require numerical inputs.

7. **Dimensionality Reduction**:
    - Dimensionality reduction techniques like Principal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) are used to reduce the number of variables while preserving important information and patterns.
8. **Dealing with Outliers:**
    - Outliers are extreme values that differ significantly from the rest of the data. Data transformation can involve handling or removing outliers to avoid undue influence on the analysis.

Data Transformation Techniques

There are several data transformation techniques that are used to clean data and structure it before it is stored in a data warehouse or analyzed for business intelligence. Not all of these techniques work with all types of data, and sometimes more than one technique may be applied. Nine of the most common techniques are:

1. Revising

Revising ensures the data supports its intended use by organizing it in the required and correct way. It does this in a range of ways.

- Dataset normalization revises data by eliminating redundancies in the data set. The data model becomes more precise and legible while also occupying less space. This process, however, does involve a lot of critical thinking, investigation and reverse engineering.
- Data cleansing ensures the formatting capability of data.
- Format conversion changes the data types to ensure compatibility.
- Key structuring converts values with built-in meanings to generic identifiers to be used as unique keys.
- Deduplication identifies and removes duplicates.
- Data validation validates records and removes the ones that are incomplete.
- Repeated and unused columns can be removed to improve overall performance and legibility of the data set.

2. Manipulation

This involves creation of new values from existing ones or changing current data through computation. Manipulation is also used to convert unstructured data into structured data that can be used by machine learning algorithms.

- Derivation, which is cross column calculations
- Summarization that aggregates values

- Pivoting which involves converting columns values into rows and vice versa
- Sorting, ordering and indexing of data to enhance search performance
- Scaling, normalization and standardization that helps in comparing dissimilar numbers by putting them on a consistent scale
- Vectorization which helps convert non-numerical data into number arrays that are often used for machine learning applications

3. Separating

This involves dividing up the data values into its parts for granular analysis. Splitting involves dividing up a single column with several values into separate columns with each of those values. This allows for filtering on the basis of certain values.

4. Combining/ Integrating

Records from across tables and sources are combined to acquire a more holistic view of activities and functions of an organization. It couples data from multiple tables and datasets and combines records from multiple tables.

5. Data Smoothing

This process removes meaningless, noisy, or distorted data from the data set. By removing outliers, trends are most easily identified.

6. Data Aggregation

This technique gathers raw data from multiple sources and turns it into a summary form which can be used for analysis. An example is the raw data providing statistics such as averages and sums.

7. Discretization

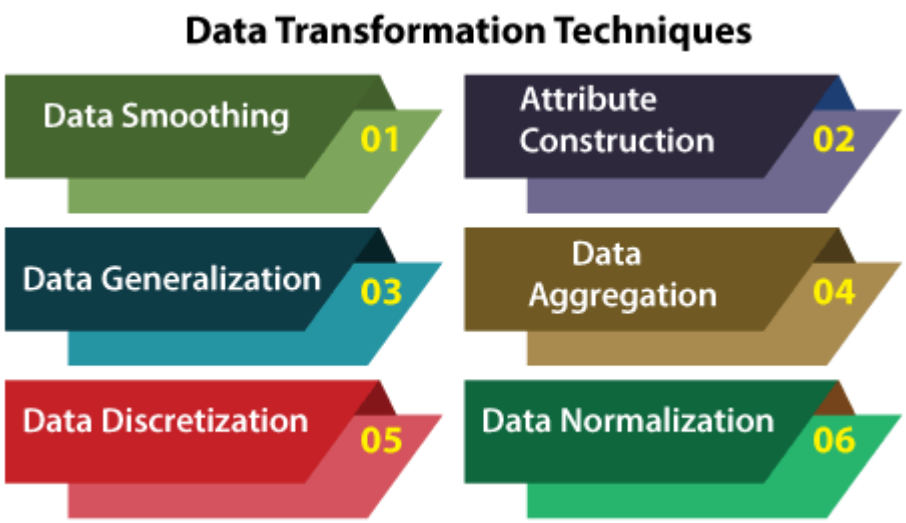
With the help of this technique, interval labels are created in continuous data in an attempt to enhance its efficiency and easier analysis. The decision tree algorithms are utilized by this process to transform large datasets into categorical data.

8. Generalization

Low level data attributes are transformed into high level attributes by using the concept of hierarchies and creating layers of successive summary data. This helps in creating clear data snapshots.

9. Attribute Construction

In this technique, a new set of attributes is created from an existing set to facilitate the mining process.


Data Transformation Techniques
Data Smoothing
01
Attribute Construction
02
Data Generalization
03
Data Aggregation
04
Data Discretization
05
Data Normalization
06