DATA MINING

DATA PREPROCESSING

1) What is Knowledge Discovery in Databases (KDD) Process?

Ans) Knowledge Discovery in Databases (KDD) Process:

A method of finding, transforming, and refining meaningful data and patterns from a raw database in order to be utilized in different domains or applications.

2) Write the Steps in typical KDD Process.

Ans) Steps in typical KDD Process:

(i) Goal-Setting and Application Understanding: Requires prior understanding and knowledge of the field to be applied in.

This is where we decide how the transformed data and the patterns arrived at by data mining will be used to extract knowledge.

(ii) Data Selection and Integration: The data collected needs to be selected and segregated into meaningful sets based on availability, accessibility importance and quality.

(iii) Data Cleaning and Preprocessing: It involves searching for missing data and removing noisy, redundant and low-quality data from the data set in order to improve the reliability of the data and its effectiveness.

Certain algorithms are used for searching and eliminating unwanted data based on attributes specific to the application.

(iv) Data Transformation: This step prepares the data to be fed to the data mining algorithms.

Hence, the data needs to be in consolidated and aggregate forms. The data is consolidated on the basis of functions, attributes, features etc.

(v) Data Mining: This is the root or backbone process of the whole KDD.

This is where algorithms are used to extract meaningful patterns from the transformed data, which help in prediction models.

It is an analytical tool which helps in discovering trends from a data set using techniques such as artificial intelligence, advanced numerical and statistical methods and specialized algorithms.

(vi) Pattern Evaluation/Interpretation: Once the trend and patterns have been obtained from various data mining methods and iterations, these patterns need to be represented in discrete forms such as bar graphs, pie charts, etc. to study the impact of data collected and transformed during previous steps.

This also helps in evaluating the effectiveness of a particular data model in view of the domain.

3) What is Data Cleaning? Explain.

Ans) DATA CLEANING: Defines to clean the data by filling in the missing values, smoothing noisy data, analyzing and removing outliers, and removing inconsistencies in the data.

Sometimes data at multiple levels of detail can be different from what is required.

Ex: It can need the age ranges of 20-30, 30-40, 40-50, and the imported data includes birth date. The data can be cleaned by splitting the data into appropriate types.

4) Types of Data Cleaning.

Ans) Types of data cleaning:

(i) Missing Values: Missing values are filled with appropriate values. There are the following approaches to fill the values.

The tuple is ignored when it includes several attributes with missing values.

The values are filled manually for the missing value.

The same global constant can fill the values.

The attribute mean can fill the missing values.

The most probable value can fill the missing values.

(ii) Noisy Data: Noise is a random error or variance in a measured variable.

(iii) Smoothing methods to handle noise:

- **Binning – These methods smooth out a arrange data value by consulting its "neighborhood," especially, the values around the noisy information.**
- **The arranged values are distributed into multiple buckets or bins because binning methods consult the neighborhood of values, they implement local smoothing.**

(iv) Regression:

- **Data can be smoothed by fitting the information to a function, including with regression.**
- **Linear regression: Contains finding the "best" line to fit two attributes (or variables) so that one attribute can be used to forecast the other.**
- **Multiple linear regression: A development of linear regression, where more than two attributes are contained and the data are fit to a multidimensional area.**

(v) Clustering:

- **Clustering supports in identifying the outliers.**
- **The same values are organized into clusters and those values which fall outside the cluster are known as outliers.**

5) TERMS RELATED TO DATA.

Ans) TERMS RELATED TO DATA:

- **ATTRIBUTE: Property or characteristic of an object. Also called as variable, field, characteristic or feature. Ex: Eye color of a person, temperature, etc.**
- **INSTANCE: Collection of attributes describing an object. Also known as record, point, case, sample or entity.**

6) TYPES OF ATTRIBUTES.

Ans) TYPES OF ATTRIBUTES:

(i) Nominal: This type of data is also referred to as categorical data that cannot be measured or compared with numbers. Ex: Gender, race, religion, and occupation.

(ii) Ordinal: Qualitative data that can be ranked in a particular order. Ex: Ranks (scale of 0 to 10), Grades, Height (tall, short, medium), etc.

(iii) Interval: Quantitative data with equal intervals between consecutive values. Ex: Time, dates, temperature, etc.

(iv) Ratio: Similar to interval data, but with an absolute zero point. Ex: height, weight, and income.

7) Properties of Attribute Values.

Ans) Properties of Attribute Values:

Type of an attribute depends on which of the following properties it

possesses:

Distinctness: =, ≠

Order: <, >

Addition: +, -

Multiplication: *

Nominal Attribute: Distinctness.

Ordinal Attribute: Distinctness and Order.

Interval Attribute: Distinctness, Order and Addition.

Ratio Attribute: All 4 attributes.

8) Discrete and Continuous Attributes.

Ans) Discrete and Continuous Attributes:

- **Discrete: Number of words in a given sentence.**
- **Continuous: Temperature, height, etc.**
- **We an convert continuous to discrete by dividing the continuous values in intervals. This is called Discretization.**