23/04/24

ABHIYANK YADAV

22MIP10074

DATA MINING AND WAREHOUSING ASSIGNMENT

(1) Data Cleaning

Data cleaning, also referred as data cleansing, is an essential processing step in the data analysis pipeline. It involves identifying and correcting errors, inconsistencies and inaccuracies in a dataset to ensure its quality and reliability for subsequent analysis.

Importance of Data Cleaning:-

(1) **Ensures Data quality**:- Cleaning data helps maintain the accuracy, completeness and consistency of the dataset.

(2) **Reduce errors and Bias**:- Cleaning data minimizes errors, biases and discrepancies that may arise for inaccurate or incomplete info.

(3) **Enhances Analysis efficiency**:- Clean data streamlines the analysis process by reducing the time and effort required to identify and address data issues during analysis.

(4) **Supports Decision Making**:- High quality Data enables decision-makers to make informed and confident decisions based on trustworthy information.

(2) Data Visualization:-

It involves representing data in a graphical or pictorial format to facilitate understanding, exploration and communication of insights. The primary goal of data visualization is to make complex dataset more accessible, intuitive, and actionable for decision-makers.

Importance of Data Visualization

① Facilitates Understanding:- Visual representation such as charts, graphs and maps make it easier for users to comprehend complex datasets.

② Reveals Patterns and trends:- Visualization allows users to identify patterns, trends, correlations and outliers.

③ Data Integration

It is the process of combining data from different sources or formats into a unified view, a single database, application or platform. The goal of data integration is to provide users with a comprehensive and consistent view of data.

~~Infor~~ Importance of Data Integration

① Unified Data View:- Integrating data from diverse sources allo organizations to create a unified view of their data assets.

② Enhanced Insights:- Integrated data enables deeper insights and analysis by providing a holistic view of business processes, customer interaction.

③ Streamlined Operation:- Data integration streamlines business operations by eliminating silos and redundancies.

④ Data Reduction

It is the process in data analysis and management aimed at reducing the volume or dimensionality of a dataset while preserving its essential charactastics and minimizing information loss. The goal of data reduction is to simplify complex datasets to make them more mangeable.

Importance of Data Reduction

① Improved Efficiency:- Reducing the size or dimensionality of a dataset can improve computational efficiency and reduce storage requirements.

② Faster Processing- Smaller datasets are typically faster to process, enambling quicker data analysis, modelling and visualization tasks.

③ Simplified Analysis:- Data reduction techniques help simplify data analysis by focusing on the most relevant or informative features.

⑤ Data Tranformation

It is a fundamental process in data preparation and analysis, involving the conversion or manipulation of data from one format, structure or representation to another. The primary objective of data tranformation is to prepare data for downstream analysis, modelling and visualization.

Importance of Data Tranformation:-

① Data Standardization:- Tranformation helps standardize data formats, units and representation across diverse sources.

② Feature Engineering:- Tranformation enables feature engineering, where new features or variables are derived or extracted from existing data.

③ Data Integration:- Tranformation plays a key role in Data Integration by harmonizing and aligning data from disparate sources to create a unified data.

⑥ Data Discretization

It is a data preprocessing technique used to reduce the number of values in a continuous dataset by partitioning the range of values into a finite number of intervals or bins. This process converts continuous data into Categorical or ordinal data, where each interval represents a distinct categories or value range.

Importance of Data Discretization

① Simplification of Analysis:- Discretization simplifies the analysis of continuous data by converting it into a discrete form.

② Reduction of noise:- Discretization can help reduce noise and variability in continuous data by grouping similar values into the same category or interval.

③ Interpretability:- Discretized data is often more interpretable than continuous data, as it provides clear boundaries and categories.