UNIT-2

Data Pre-processing:

Introduction, Data Quality, Data Cleaning- Missing Values, Noisy data, Data Integration, Data Transformation- Smoothing, Attribute construction, Aggregation, Normalization, Discretization, Data Reduction- Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Histograms, Clustering, Sampling

Data Pre-processing:

Data pre-processing is an essential step in data analysis and machine learning that involves transforming raw data into a format suitable for analysis. It involves several tasks, including data cleaning, data transformation, data integration, and data reduction, which aim to improve the quality of data for analysis. These tasks ensure that the data is properly formatted, complete, accurate, and consistent. In most cases, raw data obtained from various sources may contain errors, inconsistencies, missing values, or other issues that can adversely affect the quality of results obtained from data analysis.

Data cleaning involves identifying and correcting errors and inconsistencies in the data. This can include removing duplicate records, correcting spelling mistakes, and fixing formatting issues. Data transformation involves converting the data into a more useful format, such as scaling numerical variables, converting categorical variables into numerical ones, and handling missing values.

Data integration involves combining data from multiple sources into a single dataset for analysis. This can be challenging when the data comes from different formats or contains different types of variables. Data reduction involves reducing the dataset size without sacrificing too much information. This can be done through techniques such as sampling or principal component analysis.

Overall, data pre-processing is a crucial step in data analysis and machine learning that ensures the data used in the analysis is accurate, consistent, and properly formatted. Proper data pre-processing can improve the quality of results obtained from data analysis, leading to better decision-making and improved performance in machine learning models.

Data Quality:

Data quality refers to the accuracy, completeness, consistency, and timeliness of data, which are important factors in ensuring that the data is reliable and useful for analysis. Poor data quality can lead to incorrect conclusions, inaccurate predictions, and poor decision-making.

Here are some key factors that contribute to data quality:

1. Accuracy: The data must be correct and free from errors. This includes ensuring that numerical values are accurate, and that any text or categorical data is correctly labeled and spelled.
2. Completeness: The data must be complete, meaning that it contains all the necessary information for the analysis to be accurate. This includes ensuring that there are no missing values, and that all records are included.
3. Consistency: The data must be consistent, meaning it is uniform and follows the same rules across all records. For example, dates should be formatted the same way, and units of measurement should be consistent.
4. Timeliness: The data must be up-to-date and relevant for the analysis. This means that data should be updated regularly, and any outdated or irrelevant data should be removed.
5. Relevance: The data must be relevant to the analysis being conducted. This includes ensuring that the data is appropriate for the intended use and collected from a reliable source.

To improve data quality, several strategies can be employed, including:

1. Data profiling: This involves analyzing the data to identify any issues or inconsistencies. Data profiling can help to identify issues such as missing values, inconsistencies, and outliers.
2. Data cleansing: This involves correcting or removing any errors, inconsistencies, or inaccuracies in the data. This can include correcting spelling mistakes, removing duplicates, and filling in missing values.
3. Data enrichment: This involves adding additional data to the dataset to improve its quality. This can include adding attributes to records, or appending data from other sources.
4. Data governance: This involves implementing policies, procedures, and controls to ensure that the data is of high quality. This can include establishing data quality standards and implementing data validation and verification processes.

Overall, data quality is essential for ensuring that the data used in analysis is accurate and reliable. By improving data quality, organizations can make better decisions and improve the performance of machine learning models.

Data Cleaning- Missing Values, Noisy data:

Data cleaning is identifying and correcting dataset errors, inconsistencies, and inaccuracies. It involves several tasks, including handling missing values and removing noisy data. Here is a detailed explanation of these two tasks:

1. **Handling Missing Values:**

Missing values are values in a dataset that are not present or cannot be obtained. Handling missing values is a critical step in data cleaning as they can cause problems in statistical analysis or machine learning algorithms. There are several strategies for handling missing values, including:

a. Deleting: In some cases, deleting records or variables with missing values may be appropriate. This strategy is best used when the missing values are few and do not significantly impact the analysis.

b. Imputation: Imputation is filling in missing values with estimated or predicted values. There are several imputation methods, including mean imputation, median imputation, and mode imputation.

c. Advanced techniques: Advanced techniques such as Multiple Imputation or K-nearest neighbour (KNN) imputations can be used for more complex datasets.

Example:

Suppose a dataset contains the height and weight of 100 individuals, but for 5 individuals, the weight is missing. One strategy for handling missing values is to impute the missing weight values using the average weight of the remaining 95 individuals.

2. **Removing Noisy Data:**

Noisy data refers to data that contains errors, outliers, or inconsistencies that are not representative of the actual data. Removing noisy data is important because it can affect the accuracy of statistical analysis or machine learning models. There are several strategies for removing noisy data, including:

a. Z-score: This method involves calculating the Z-score for each record in the dataset and removing records with a Z-score above a certain threshold.

b. Interquartile Range (IQR): This method involves calculating the IQR for each variable and removing records with values outside the upper and lower boundaries of the IQR.

c. Visual inspection: This involves manually inspecting the data using graphs or charts and removing any outliers or errors.

Example:

Suppose a dataset contains the salaries of 100 employees, and one employee's salary is significantly higher than the others. This outlier could be considered noisy data and may be removed using one of the above methods to improve the accuracy of any analysis or modeling.

Data Integration:

Data integration is the process of combining data from different sources into a single, unified view. It involves merging data from disparate sources, resolving inconsistencies or duplicates, and creating a comprehensive view of the data. Here is a detailed explanation of data integration with examples:

Data Integration Steps:

1. Identify the data sources: The first step in data integration is to identify the data sources that need to be combined. These sources could include databases, spreadsheets, or files.
2. Extract the data: Once the data sources are identified, the data must be extracted from each source. This could involve exporting data from a database, copying data from a spreadsheet, or importing data from a file.
3. Transform the data: The data from each source may be in different formats or structures. It is essential to transform the data to ensure that it is consistent and can be merged. This step could involve data cleaning, data normalization, or data aggregation.
4. Merge the data: After the data is transformed, it can be merged into a single dataset. This involves matching records from each source based on a common attribute, such as a customer ID or a product code.

5. Resolve any inconsistencies: In some cases, the data may contain inconsistencies or duplicates. It is important to resolve these issues to ensure the accuracy of the final dataset.
6. Store the integrated data: The final step in data integration is to store the integrated data in a format that is easily accessible for analysis or reporting.

Example:

Suppose a company has customer data stored in two different databases. One database contains customer names, addresses, and phone numbers, while the other contains customer purchase history. To perform analysis on the data, it is necessary to integrate the data from both databases. Here's how data integration could be performed in this case:

1. Identify the data sources: The two data sources are the customer database and the purchase history database.
2. Extract the data: The customer data can be extracted by querying the customer database, and the purchase history can be extracted by querying the purchase history database.
3. Transform the data: The customer data and purchase history data may be in different formats or have different structures. The data can be transformed by renaming columns to ensure consistency, cleaning up any data errors, and aggregating data to match a common attribute, such as a customer ID.
4. Merge the data: After transforming the data, the customer and purchase history data can be merged based on the customer ID.
5. Resolve any inconsistencies: The data may contain inconsistencies or duplicates. For example, a customer may have multiple phone numbers or purchase records. These issues can be resolved by selecting the most recent phone number or purchase record.
6. Store the integrated data: The integrated data can be stored in a database or a file format such as CSV, making it easily accessible for analysis or reporting.

Data Transformation- Smoothing, Attribute construction, Aggregation, Normalization, Discretization:

Data transformation is the process of converting raw data into a format that is more suitable for analysis. Here are some of the most commonly used data transformation techniques:

1. Smoothing: Smoothing is a technique used to remove noise from data. This technique is often used in time-series data analysis to remove random fluctuations and reveal trends. A simple example of smoothing is a moving average, which calculates the average of a sliding window of data points.
2. Attribute construction: Attribute construction is a technique used to create new attributes or features from existing attributes. For example, a new attribute such as "total spending" could be constructed in a customer database by adding up the values of all the customer's purchases.
3. Aggregation: Aggregation is a technique used to summarize data at a higher level. For example, in a sales database, data could be aggregated by month, quarter, or year to provide a higher-level view of sales trends.
4. Normalization: Normalization is a technique to rescale data to a common range. This technique is often used in machine learning to ensure that all attributes are treated equally. For example, in a customer database, the age attribute could be normalized by scaling it to a range between 0 and 1.
5. Discretization: Discretization is a technique used to transform continuous data into discrete categories. For example, in a survey database, the age attribute could be discretized into categories such as "18-24", "25-34", "35-44", and so on.

Examples:

1. Smoothing: Suppose you have a dataset of daily stock prices for a company. The prices may fluctuate randomly due to market conditions. To reveal the long-term trend, you can use a moving average to smooth the data.
2. Attribute construction: Suppose you have a dataset of customer purchases. You could construct a new attribute called "total spending" by adding the values of all the customer's purchases.
3. Aggregation: Suppose you have a sales database with daily sales data. You could aggregate the data by month to provide a higher-level view of sales trends.
4. Normalization: Suppose you have a customer database with age, income, and spending attributes. You could normalize the data to ensure that all attributes

are treated equally. For example, you could scale the age attribute to a range between 0 and 1.

5. Discretization: Suppose you have a survey database with an age attribute. You could discretize the age attribute into categories such as "18-24", "25-34", "35-44", and so on. This would allow you to analyze the survey data by age group.

Data Reduction- Wavelet Transforms, Principal Components Analysis, Attribute Subset Selection, Histograms, Clustering, Sampling:

Data reduction reduces data's volume and complexity without losing its essential information. Here are some of the most commonly used data reduction techniques:

1. Wavelet Transforms: Wavelet transforms are a technique used to compress data by representing it in terms of a set of wavelets, which are mathematical functions that can represent data in a compact form. This technique is often used in signal processing to compress audio and video data.
2. Principal Components Analysis (PCA): PCA is a technique used to reduce the dimensionality of data by identifying the most important variables that explain the variance in the data. This technique is often used in machine learning to identify the most relevant features for modeling.
3. Attribute Subset Selection: Attribute subset selection is a technique used to select a subset of relevant attributes from a larger set of attributes. This technique is often used in machine learning to simplify models and improve performance.
4. Histograms: Histograms are a technique used to summarize data by dividing it into a set of bins and counting the number of data points in each bin. This technique is often used in data visualization to reveal patterns and trends in the data.
5. Clustering: Clustering is a technique used to group data points into clusters based on their similarity. This technique is often used in machine learning to identify patterns in the data.
6. Sampling: Sampling is a technique used to reduce the volume of data by selecting a representative subset of the data. This technique is often used in data analysis to speed up processing and reduce storage requirements.

Examples:

1. Wavelet Transforms: Suppose you have a large dataset of audio data. You can use wavelet transforms to compress the data by representing it in terms of a set of wavelets. This will allow you to store and process the data more efficiently.
2. Principal Components Analysis (PCA): Suppose you have a large dataset with many variables. You can use PCA to identify the most important variables that explain the variance in the data. This will allow you to simplify your models and improve performance.
3. Attribute Subset Selection: Suppose you have a dataset with many attributes. You can use attribute subset selection to identify the most relevant attributes for modeling. This will allow you to simplify your models and improve performance.
4. Histograms: Suppose you have a dataset of customer transactions. You can use histograms to summarize the data by dividing it into a set of bins and counting the number of transactions in each bin.