DATA MINING AND WAREHOUSING

Data Mining

- Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems.
- It primarily turns raw data into useful information.
- Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective.
- such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Sources of data: -

Business –Web, E-commerce, Transactions, Stocks

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining tools

1. Orange Data Mining:

2. SAS Data Mining(**Statistical Analysis System.)**

3. DataMelt Data Mining:

4. Rattle:

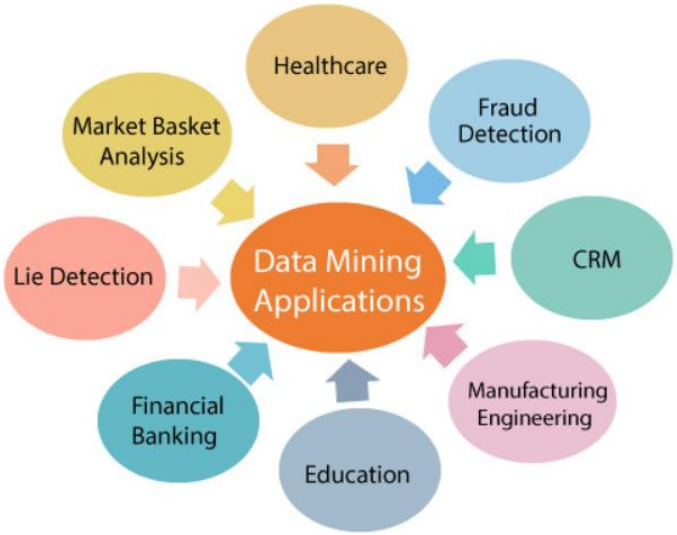
5. Rapid Miner:

Important features of Data Mining:

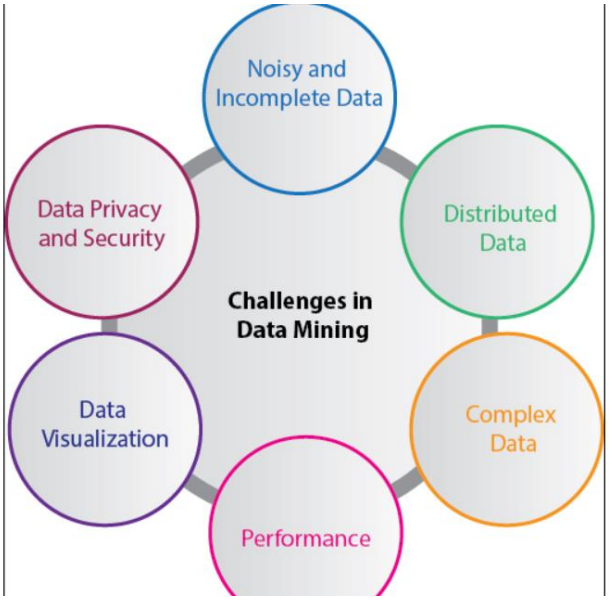
The important features of Data Mining are given below:

- It utilizes the Automated discovery of patterns.
- It predicts the expected results.
- It focuses on large data sets and databases
- It creates actionable information

Data Mining Applications



Challenges of Implementation in Data mining



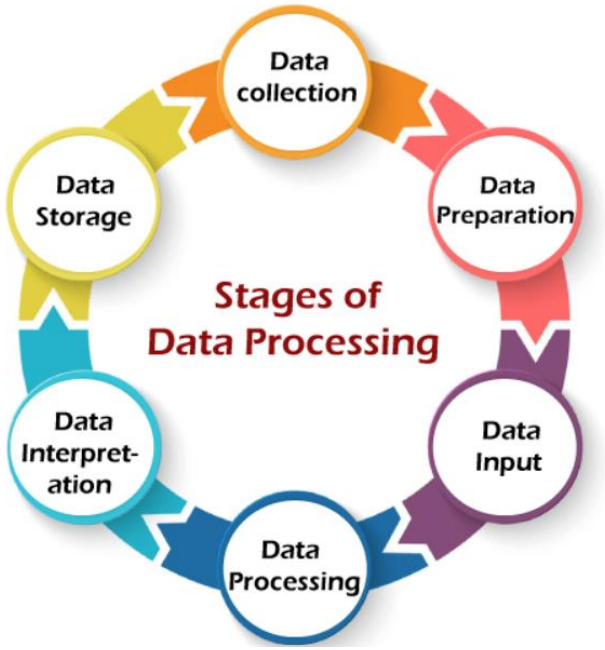
Data Mining	Data Warehousing
Discover patterns and insights	Store and manage data
Extraction of hidden information	Efficient data storage
Involves data preprocessing	Data extraction and loading
Decision-making and prediction	Reporting and analysis
Utilizes clustering, regression	Data integration and indexing
Handles historical and current	Historical data for analysis
Deals with subsets of data	Manages large data quantities
Targeted at analysts, scientists	Serves business users
Outputs patterns and predictions	Produces reports and dashboards
Analyzes fine-grained data	Focuses on aggregated data
May require data integration	Integrates data from sources
Supports real-time or batch	Primarily used for batch
Applied in predictive analytics	Business intelligence
Addresses data quality	Emphasizes data accuracy
Example: Recommender systems	Example: Sales reporting

Data Processing

- Data processing is collecting raw data and translating it into usable information.
- The raw data is ***collected, filtered, sorted, processed, analyzed, stored***, and then presented in a readable format.
- It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.

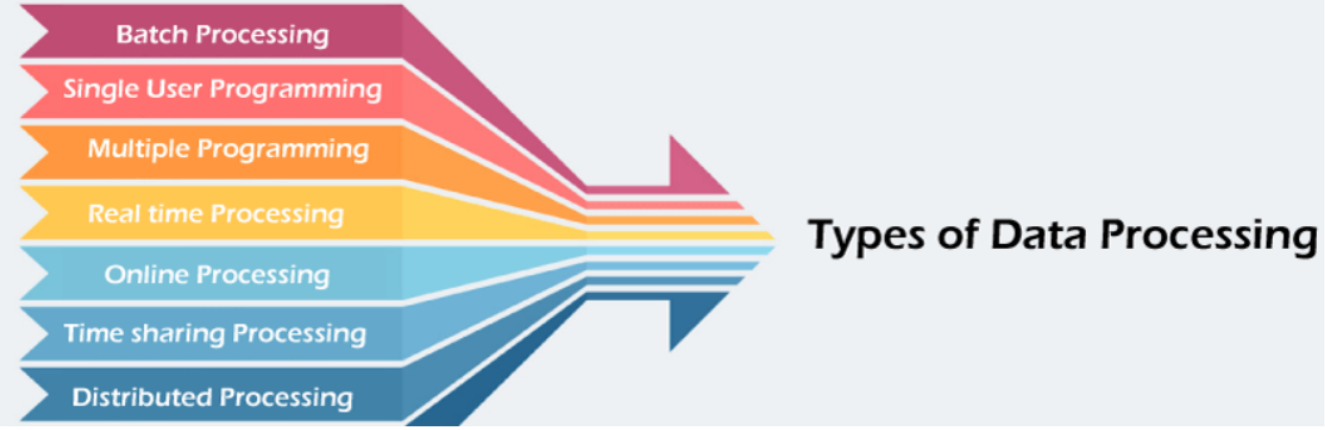
Stages of Data Processing

The data processing consists of the following six stages.



Types of Data Processing

There are different types of data processing based on the source of data and the steps taken by the processing unit to generate an output. There is no one size fits all method that can be used for processing raw data.



Examples of Data Processing

Data processing occurs in our daily lives whether we may be aware of it or not. Here are some real-life examples of data processing, such as:

- Stock trading software that converts millions of stock data into a simple graph.
- An e-commerce company uses the search history of customers to recommend similar products.
- A digital marketing company uses demographic data of people to strategize location-specific campaigns.
- A self-driving car uses real-time data from sensors to detect if there are pedestrians and other cars on the road.

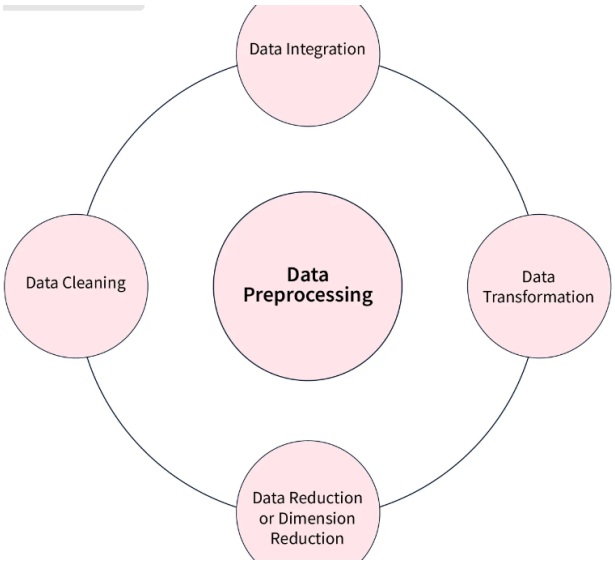
Data Preprocessing

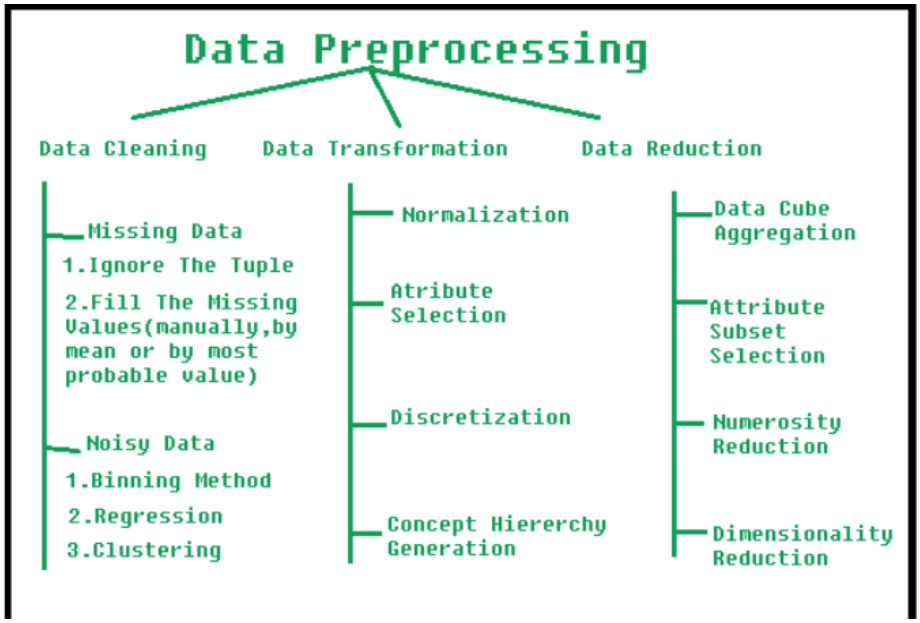
- Real-world datasets are generally messy, raw, incomplete, inconsistent, and unusable.
- It can contain manual entry errors, missing values, inconsistent schema, etc.
- Data Preprocessing is the process of converting raw data into a format that is understandable and usable.
- It is a crucial step in any Data Science project to carry out an efficient and accurate analysis.
- It ensures that data quality is consistent before applying any **Machine Learning** or **Data Mining techniques**.

Why is Data Preprocessing Important

- **Accuracy**
- **Completeness**
- **Consistent**
- **Timeliness**
- **Trustable**
- **Interpretability**

Steps in Data Preprocessing





Applications of Data Preprocessing

- **Improved Accuracy of ML Models**
- **Reduced Costs**
- **Visualization**

Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  - **Ignore the tuples:**
    This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

  - **Fill the Missing values:**
    There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**
  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
    - **Binning Method:**
      This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

    - **Regression:**
      Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

    - **Clustering:**
      This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in

hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

Sampling: This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

- # DATA INTEGRATION:

Data integration is one of the steps of data pre-processing that involves combining data residing in different sources and providing users with a unified view of these data.

• It merges the data from multiple data stores (data sources)

• It includes multiple databases, data cubes or flat files.

• Metadata, Correlation analysis, data conflict detection, and resolution of semantic heterogeneity contribute towards smooth data integration.

• There are mainly 2 major approaches for data integration - commonly known as "tight coupling approach" and "loose coupling approach".

Tight Coupling

o Here data is pulled over from different sources into a single physical location through the process of ETL - Extraction, Transformation and Loading.

o The single physical location provides an uniform interface for querying the data.

ETL layer helps to map the data from the sources so as to provide a uniform data

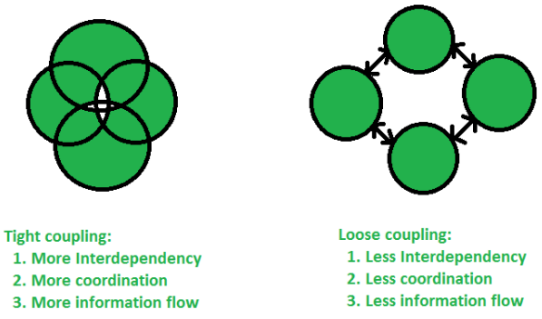
o warehouse. This approach is called tight coupling since in this approach the data is tightly coupled with the physical repository at the time of query.

ADVANTAGES:

1. Independence (Lesser dependency to source systems since data is physically copied over)
2. Faster query processing
3. Complex query processing
4. Advanced data summarization and storage possible
5. High Volume data processing

DISADVANTAGES: 1. Latency (since data needs to be loaded using ETL)

6. Costlier (data localization, infrastructure, security)



Difference between tight coupling and loose coupling

- Tight coupling is not good at the test-ability. But loose coupling improves the test ability.
- Tight coupling does not provide the concept of interface. But loose coupling helps us follow the GOF principle of program to interfaces, not implementations.
- In Tight coupling, it is not easy to swap the codes between two classes. But it's much easier to swap other pieces of code/modules/objects/components in loose coupling.
- Tight coupling does not have the changing capability. But loose coupling is highly changeable.

Data Redundancy :

It is defined as the redundancy means duplicate data and it is also stated that the same parts of data exist in multiple locations into the database. This condition is known as Data Redundancy.

Problems with Data Redundancy :

Here, we will discuss the few problems with data redundancy as follows.

1. Wasted Storage Space.
2. More Difficult Database Update.
3. It will lead to Data Inconsistency.
4. Retrieval of data is slow and inefficient.

Example –

Let us take an example of a cricket player table.

Aspect	Data Inconsistency	Data Reduction
Definition	Discrepancies in data values	Reducing data volume
Issue	Conflicting or inaccurate data	Simplifying data
Causes	Errors, outdated info, integration issues	Aggregation, filtering, simplification
Impact	Incorrect analysis, decisions	Improved processing efficiency
Examples	Differing prices for the same product	Monthly sales totals from daily data
Data Quality Improvement	Data cleansing, reconciliation	Streamlining for efficiency
Goal	Ensure data accuracy	Enhance processing efficiency
Nature of Process	Error correction	Data simplification
Use Case	Data quality assurance	Performance optimization
End Result	Consistent, accurate data	Streamlined, manageable

Regenerate

Data Transformation in Data Mining

- Raw data is difficult to trace or understand.
- That's why it needs to be preprocessed before retrieving any information from it.
- Data transformation is a technique used to **convert** the raw data into a suitable format that efficiently eases data mining and retrieves strategic information.

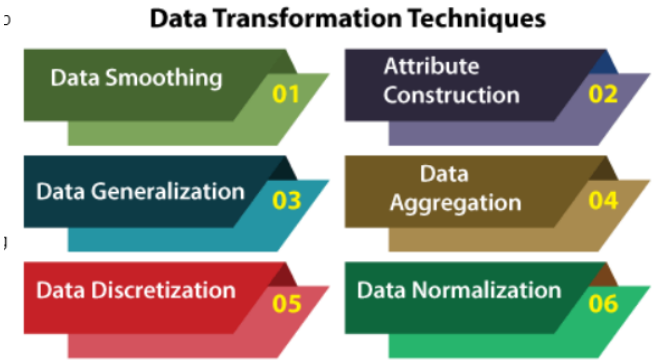
➤ Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form.

- Data transformation is an essential data preprocessing technique that must be performed on the data before data mining to provide patterns that are easier to understand.
- Data transformation changes the format, structure, or values of the data and converts them into clean, usable data.

- **Constructive:** The data transformation process adds, copies, or replicates data.
- **Destructive:** The system deletes fields or records.
- **Aesthetic:** The transformation standardizes the data to meet requirements or parameters.
- **Structural:** The database is reorganized by renaming, moving, or combining columns.

Data Transformation Techniques

There are several data transformation techniques that can help structure and clean up the data before analysis or storage in a data warehouse. Let's study all techniques used for data transformation, some of which we have already studied in data reduction and data cleaning.



Data Cube Aggregation(OLAP)

Online Analytical Processing (OLAP)

This technique is used to aggregate data in a simpler form.

Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction.

For example, suppose you have the data of All Electronics sales per quarter for the year 2018 to the year 2022. If you want to get the annual sale per year, you just have to aggregate the sales per quarter for each year. In this way, aggregation provides you

with the required data, which is much smaller in size, and thereby we achieve data reduction even without losing any data.

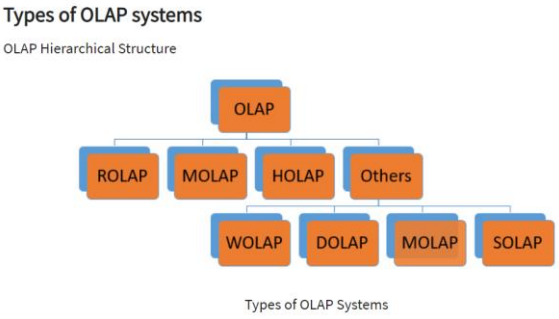
What Are the Data Cube Operations?

Basic analytical operations of OLAP

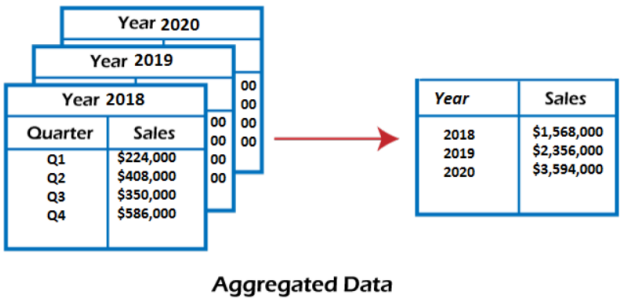
Four types of analytical OLAP operations are:

1. Roll-up
2. Drill-down
3. Slice and dice
4. Pivot (rotate)

- **Rollup** – decreases dimensionality by aggregating data along a certain dimension
- **Drill-down** – increases dimensionality by splitting the data further
- **Slicing** – decreases dimensionality by choosing a single value from a particular dimension
- **Dicing** – picks a subset of values from each dimension
- **Pivoting** – rotates the data cube



Types of OLAP Systems



Aggregated Data

The data cube aggregation is a multidimensional aggregation that eases multidimensional analysis. The data cube present precomputed and summarized data which eases the data mining into fast access.

Advantages of Data Cube

1. Faster analysis

2. More informative visualizations

3. Intuitive navigation in large datasets

4. Faster query processing

5. Easier sharing with colleagues

Dimensionality Reduction

It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information.

such as

- **speech recognition,**
- **signal processing,**
- **bioinformatics, etc.**
- **It can also be used for data visualization, noise reduction, cluster analysis**, etc

Components of Dimensionality Reduction

There are two components of dimensionality reduction:

- **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
  1. Filter
  2. Wrapper
  3. Embedded

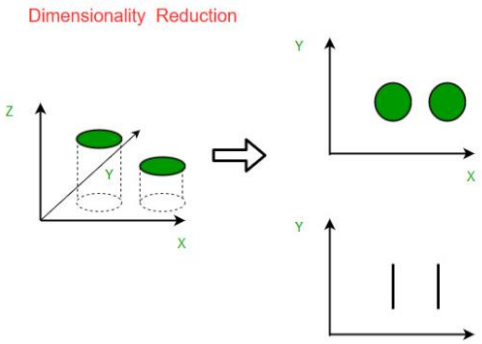
- **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Methods of Dimensionality Reduction

The various methods used for dimensionality reduction include:

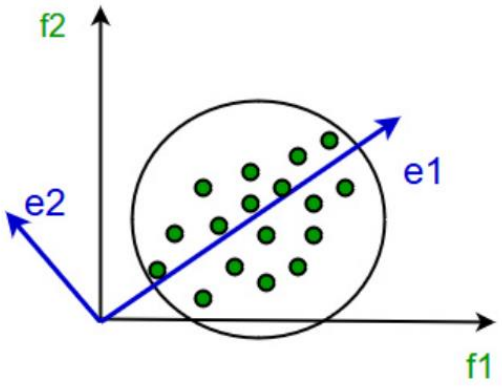
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA)

Dimensionality reduction may be both linear and non-linear, depending upon the method used. The prime linear method, called Principal Component Analysis, or PCA, is discussed below.



Principal Component Analysis

- This method was introduced by Karl Pearson.
- It works on the condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.



It involves the following steps:

- Construct the covariance matrix of the data.
- Compute the eigenvectors of this matrix.
- Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.
- Improved Visualization:

Disadvantages of Dimensionality Reduction

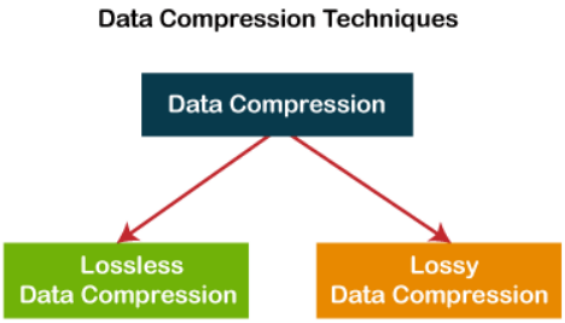
- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.

Data Compression:

The data compression technique reduces the size of the files.

- The advantage of data compression is that it helps us save our disk space and time in the data transmission.
- There are mainly two types of data compression techniques –

1. Lossless Data Compression
2. Lossy Data Compression



1. Lossless data compression

- It is used to compress the files **without losing an original file's quality and data**.
- lossless data compression, file size is reduced, but the quality of data remains the same.
- The main advantage of lossless data compression is that we can restore the original data in its original form after the decompression.
- Lossless data compression mainly used in documents, confidential information, **and PNG, RAW, GIF, BMP file formats.**

Lossless data compression techniques are -

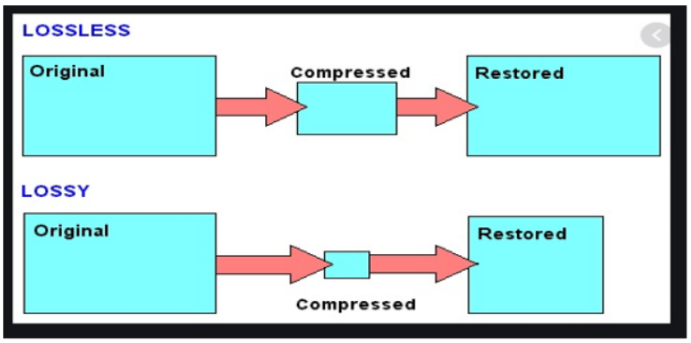
1. Run Length Encoding (RLE)
2. Lempel Ziv - Welch (LZW)
3. Huffman Coding
4. Arithmetic Coding

2. Lossy data compression

- Lossy data compression is used to compress larger files into smaller files.
- In this compression technique, some specific amount of **data and quality are removed (loss) from the original file**.
- It takes less memory space from the original file due to the loss of original data and quality.
- This technique is generally useful for us when the quality of data is not our first priority.
- Lossy data compression is most widely used in **JPEG images, MPEG video, and MP3 audio formats.**

Lossy data compression techniques are -

1. Transform coding
2. Discrete Cosine Transform (DCT)
3. Discrete Wavelet Transform (DWT)



Advantages of data compression

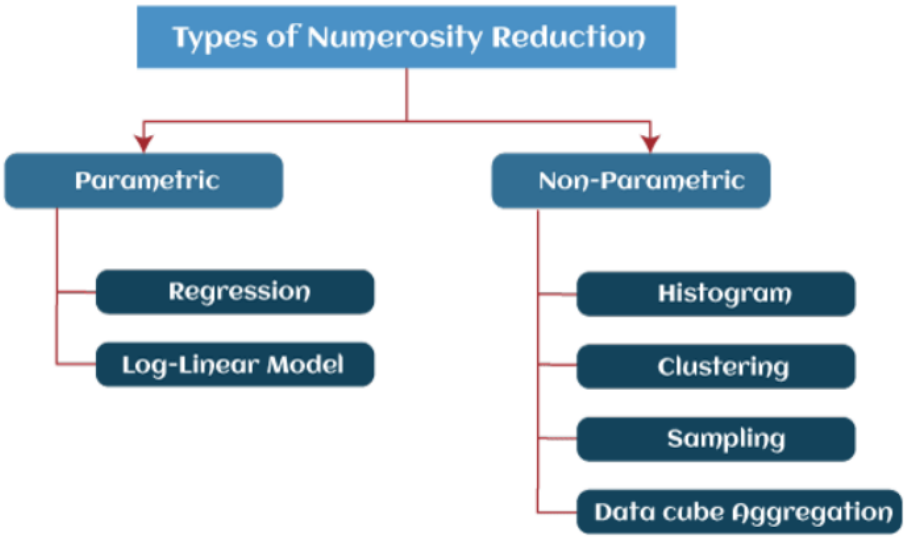
- Faster reading and writing.
- It occupies less storage space.
- File transmission takes place faster.

Disadvantages of data compression

- Added with complications.
- Effect of transmission errors.

Numerosity Reduction

- Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation.
- This can be beneficial in situations where the dataset is too large to be processed efficiently.
- There are two techniques for numerosity reduction- **Parametric** and **Non-Parametric** methods.



Parametric Methods –

❖ **Regression:**

- Regression can be a simple linear regression or multiple linear regression.
- In linear regression, the data are modeled to a fit straight line.
- y = ax+b where a and b (regression coefficients) specifies the slope and y-intercept of the line,

❖ **Log-Linear Model:**

- This allows a higher-dimensional data space to be constructed from lower-dimensional attributes.
- *Regression and log-linear model can both be used on sparse data, although their application may be limited.*

Non-Parametric Methods –

- **Histograms:** Histogram is the data representation in terms of frequency.
- **Clustering:** Clustering divides the data into groups/clusters.
- **Sampling:** Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset).
- **Data Cube Aggregation:** Data cube aggregation involves moving the data from detailed level to a fewer number of dimensions.

Advantages:

- **Improved efficiency**:
- **Improved performance:**
- **Reduced storage costs:**
- **Improved interpretability:**

Disadvantages:

- **Loss of information**
- **Impact on accuracy:**
- **Impact on interpretability**:
- **Additional computational costs**:

5. Discretization Operation

- Discretization is one form of data transformation technique.
- It transforms numeric values to interval labels of conceptual labels.
- Ex. age can be transformed to (0-10,11-20....) or to conceptual labels like youth, adult, senior.

i. **Top-down discretization:**
ii. **Bottom-up discretization:**

There are different **techniques of discretization:**

1. **Discretization by binning:** It is unsupervised method of partitioning the data based on equal partitions , either by equal width or by equal frequency.

2. **Discretization by Cluster:** clustering can be applied to discretize numeric attributes. It partitions the values into different clusters or groups by following top down or bottom up strategy.
3. **Discretization By decision tree:** it employs top down splitting strategy. It is a supervised technique that uses class information.
4. **Discretization By correlation analysis:** ChiMerge employs a bottom-up approach by finding the best neighboring intervals and then merging them to form larger intervals, recursively
5. **Discretization by histogram:** Histogram analysis is unsupervised learning because it doesn't use any class information like binning. There are various partition rules used to define histograms.

Discretization By Histogram:

Histogram analysis is unsupervised learning because it doesn't use any class information like binning.

Example: The following data shows the price of commonly sold items in sorted order: 1,1,4,4,4,4,7,7,9,9,9,9,9,11, 13,13,13,17,17,17,17,17,17, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30,30, 30.

Following figure shows histogram for the current data:

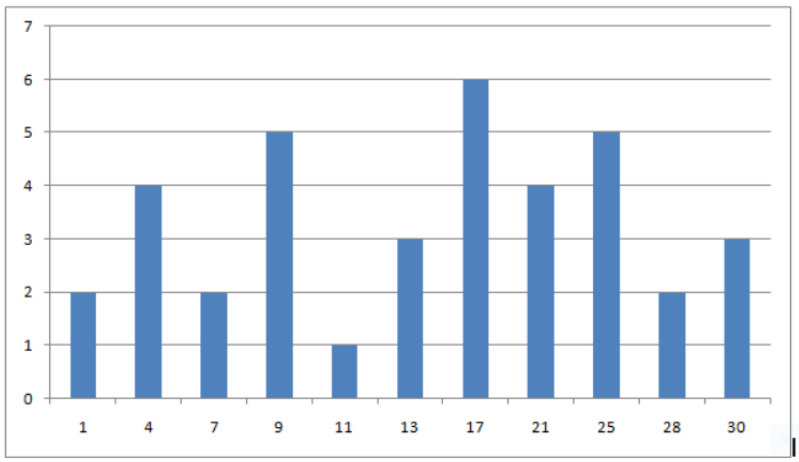


Figure 1 Histogram using price where one bucket represents one value

Now, we will partition into equal width bins where every bucket has same size width of 10.

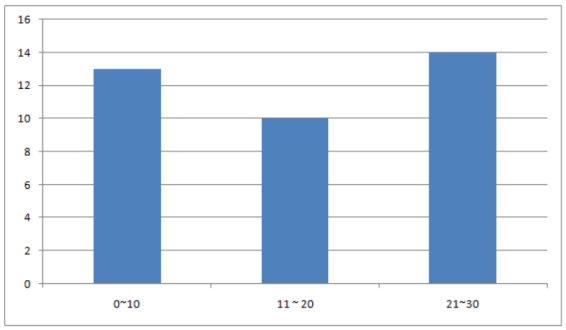


Figure 2 : Equal width Histogram

Importance of Discretization:

A discretization is important because it is useful:

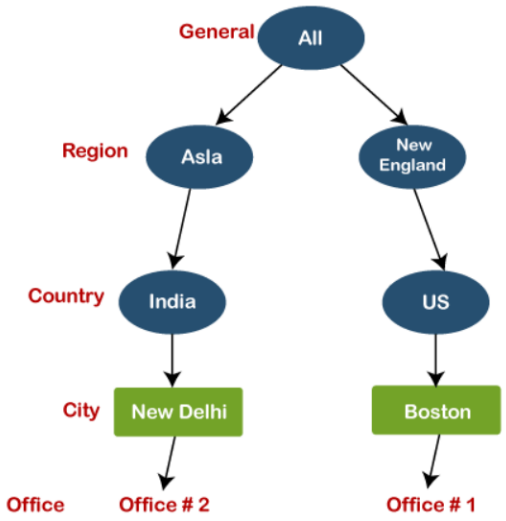
1. To generate concept hierarchies.
2. Transform numeric data.
3. To ease evaluation and management of data.
4. To minimize data loss.
5. To produce a better result.
6. Generate a more understandable structure viz. decision tree.

Concept Hierarchy in Data Mining

- the concept of a concept hierarchy refers to the organization of data into a tree-like structure,
- where each level of the hierarchy represents a concept that is more general than the level below it.
- This hierarchical organization of data allows for more efficient and effective data analysis,
- The concept of hierarchy is used to organize and classify data in a way that makes it more understandable and easier to analyze.
- it is easier to understand and perform analysis.

Bottom-up mapping

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.



Concept Hierarchy Generation

Need of Concept Hierarchy in Data Mining

- **Improved Data Analysis**
- **Improved Data Visualization and Exploration**:
- **Improved Algorithm Performance**:
- **Data Cleaning and Pre-processing**:
- **Domain Knowledge**:

Applications of Concept Hierarchy

- **Data Warehousing**:

- **Business Intelligence**:
- **Online Retail**:
- **Healthcare**:
- **Natural Language Processing**:
- **Fraud Detection**: