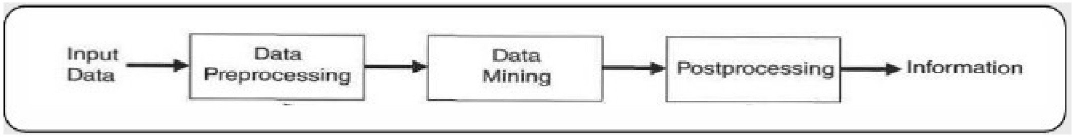
Data: refers to raw symbols or facts that represent properties of objects, events, and their environments.

Information: Data that is processed, organized, or presented in a context to make it meaningful.

Knowledge: Formulate a set of facts in rules to use them to make decision.

Data mining: is the process of extracting useful information from large volumes of data.



Pattern: is an arrangement of repeated parts. It is a summarization of relationships in the data, perhaps holding for only a few records or a few variables or both.

Model: is a global summary of data sets that can describe or summarize the unknown relations between the data items (or subset of patterns).

Classification: derives a model from classified data to determine the class of a new object based on its attributes.

Clustering: finds groups of closely related data objects in data set so that objects that belong to the same cluster (group) are more similar to each other and more dissimilar with objects in other cluster.

KDD (knowledge discover from data): is a process of detecting interesting patterns in the data It includes: data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation

Data cleaning: is the process where the data gets cleaned (remove noise and inconsistent data, and handle missing values).

Data integration: is the process where data from different data sources are integrated into one.

Data selection: is the process where the data relevant to the analysis is retrieved from the database (data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations).

Data transformation: is the process of transforming and consolidating the data into different forms that are suitable for mining

pattern evaluation: identifies the truly interesting patterns representing knowledge based on different types of interestingness measures.

Data Set: is a collection of data objects and their attributes

Attribute: is a property or characteristic of an object.

Attribute: is also known as variable, field, characteristic, dimension, or feature .

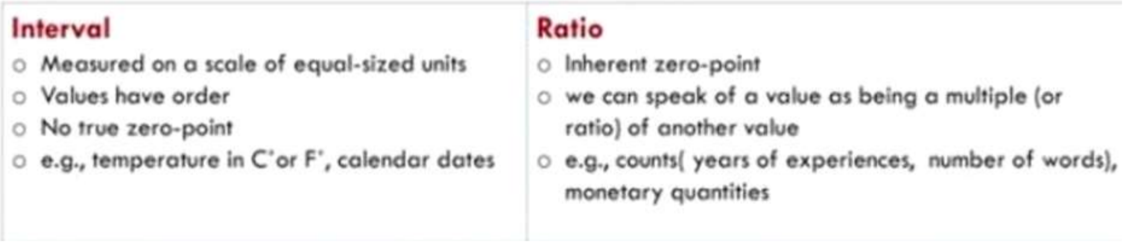
Object: represents an entity where a collection of attributes describe it.

Object: is also known as record, point, case, sample, entity, or instance .

Nominal: categories, states, or names of thing .

Ordinal: values have meaningful order(ranking) but the magnitude between two successive values is not known.

Binary: nominal attribute with only two values .



Interval

- Measured on a scale of equal-sized units
- Values have order
- No true zero-point
- e.g., temperature in C˚or F˚, calendar dates

Ratio

- Inherent zero-point
- we can speak of a value as being a multiple (or ratio) of another value
- e.g., counts(years of experiences, number of words), monetary quantities

Discrete attribute: has only finite or countable infinite set of values Sometimes the attribute represented as an integer variables.

Binary attribute: special case of discrete attributes.

Continuous attributes: has real numbers as attribute values Practically, real values can only be measured and represented as floating point variables.

Similarity Measure: Numerical measure of how alike two objects are. Values are higher when objects are more alike. Often falls in the range [0, 1].

Dissimilarity Measure: Numerical measure of how different two objects are. Lower when objects are more alike. Minimum dissimilarity is often 0. Upper limit varies.

Proximity: refers to similarity or dissimilarity.

Accuracy: correct or wrong, accurate or not (garbage in -> garbage out)

Completeness: not recorded, unavailable

Consistency: inconsistent naming, coding, format, values

Timeliness: timely updated

Believability: how trustable the data are correct

Interpretability: how easy the data are understood .

Linear regression: involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.

outliers: values that fall outside of the set of clusters.

From exams:-

Entity Identification Problem: The challenge of correctly identifying and merging records that refer to the same real-world entity in a dataset.

Attribute Redundancy: Occurs when multiple attributes in a dataset provide the same or very similar information, leading to inefficiency.

Irrelevant Attributes: Attributes that do not contribute to the analysis or prediction of the target variable and can be removed from the dataset.

Data Integration: The process of combining data from different sources to create a unified dataset for analysis.

Data Reduction: Techniques used to reduce the size of the dataset while retaining important information, such as dimensionality reduction.

Normalization: The process of scaling numerical data to a common range, usually between 0 and 1, to improve the performance of machine learning models.

Pattern Evaluation: The process of assessing and validating the discovered patterns or models based on their usefulness and relevance to the problem.

Pattern Presentation: The method of displaying or communicating discovered patterns in an understandable and actionable form.

Discrete Attribute: An attribute that has a finite number of distinct values or categories, such as a binary or nominal variable.

Redundant Attribute: An attribute that provides duplicate or highly correlated information with another attribute in the dataset.

Temporal Data: Data that represents information about events or observations over time, such as timestamps or time-related values.

Time Series Data: A sequence of data points collected or recorded at regular time intervals, often used for trend analysis or forecasting.

Spatial Data: Data that represents information about objects in physical space, including geographical coordinates and locations.

Normal Distribution Data: Data that follows a bell-shaped curve where most values are clustered around the mean, with a symmetric distribution.

Positively Skewed Data: Data where the tail on the right side is longer than on the left, indicating that most values are lower but a few are extremely high.

Negatively Skewed Data: Data where the tail on the left side is longer than on the right, indicating that most values are higher but a few are extremely low.

Noise Data: Unwanted or irrelevant data that can obscure the true patterns or relationships in a dataset, often caused by errors or inconsistencies.

Missing Data: Data that is absent or unrecorded for certain observations or attributes in a dataset, requiring imputation or removal.

Inconsistent Data: Data that contains contradictions or discrepancies, often due to errors during data entry or integration