RGPVNOTES.IN

Please do not share these notes on apps like WhatsApp or Telegram.

The revenue we generate from the ads we show on our website and app funds our services. The generated revenue **helps us prepare new notes and improve the quality of existing study materials**, which are available on our website and mobile app.

If you don't use our website and app directly, it will hurt our revenue, and we might not be able to run the services and **have to close them**. So, it is a humble request for all to **stop sharing the study material** we provide on various apps. Please **share the website's URL instead.**

Unit-III

Introduction to Data & Data Mining: Data Types, Quality of data, Data Preprocessing, Similarity measures, Summary statistics, Data distributions, Basic data mining tasks, Data Mining V/s Knowledge discovery in databases. Issues in Data mining. Introduction to Fuzzy sets and Fuzzy logic.

Introduction to Data

In computing, data is information that has been translated into a form that is efficient for movement or processing. Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural subject. Raw data is a term used to describe data in its most basic digital format.

Computers represent data, including video, images, sounds and text, as binary values using patterns of just two numbers: 1 and 0. A bit is the smallest unit of data, and represents just a single value. A byte is eight binary digits long. Storage and memory is measured in megabytes and gigabytes. Growth of the web and smartphones over the past decade led to a surge in digital data creation. Data now includes text, audio and video information, as well as log and web activity records. Much of that is unstructured data.

In general, data is any set of characters that is gathered and translated for some purpose, usually analysis. If data is not put into context, it doesn't do anything to a human or computer.

There are multiple types of data. Some of the more common types of data include the following:

- Single character
- Boolean (true or false)
- Text (string)
- Number (integer or floating-point)
- Picture
- Sound
- Video

In a computer's storage, data is a series of bits (binary digits) that have the value one or zero. Data is processed by the CPU, which uses logical operations to produce new data (output) from source data (input).

Introduction to Data Mining

Data Mining (DM) is processing data to identify patterns and establish relationships. DM is the process of analyzing data from different perspectives and summarizing it into useful information.

This information can be used in decision making. DM is the extraction of hidden predictive information from large amounts of data stored in the data warehouse for useful information, using technology with great potential to help companies focus on the most important information.

Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

"Data mining is the identification or extraction of relationships and patterns from data using computational algorithms to reduce, nodal, understand, or analyze data."

Data Mining Functionalities

Different kind of patterns can be discovered depending on the data mining task in use. There are mainly two types of data mining tasks:

1. Descriptive Data Mining Tasks
2. Predictive Data Mining Tasks

Descriptive mining tasks characterize the common properties of the existing data. Predictivemining tasks perform inference on the existing data in order to make predictions.

Types of data that can be mined

1. Data stored in the database

A database is also called a database management system or DBMS. Every DBMS stores data that are related to each other in a way or the other. It also has a set of software programs that are used to manage data and provide easy access to it. These software programs serve a lot of purposes, including defining structure for database, making sure that the stored information remains secured and consistent, and managing different types of data access, such as shared, distributed, and concurrent.

2. Data warehouse

A data warehouse is a single data storage location that collects data from multiple sources and then stores it in the form of a unified plan. When data is stored in a data warehouse, it undergoes cleaning, integration, loading, and refreshing. Data stored in a data warehouse is organized in several parts.

3. Transactional data

Transactional database stores records that are captured as transactions. These transactions include flight booking, customer purchase, click on a website, and others. Every transaction record has a unique ID. It also lists all those items that made it a transaction.

4. Other types of data

We have a lot of other types of data as well that are known for their structure, semantic meanings, and versatility. They are used in a lot of applications. Here are a few of those data types: data streams, engineering design data, sequence data, graph data, spatial data, multimedia data, and more.

Quality of data

Data quality is a measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability and whether it's up to date.

Data quality enables you to cleanse and manage data while making it available across your organization. High-quality data enables strategic systems to integrate all related data to provide a complete view of the organization and the interrelationships within it. Data quality is an essential characteristic that determines the reliability of decision-making.

Data is a valuable asset that must be managed as it moves through an organization. As information sources are growing more numerous and diverse, and regulatory compliance initiatives more focused, the need to integrate, access and reuse information from these disparate sources consistently and trustfully is becoming critical.

Data Preprocessing

1. Real world data are generally:

   Incomplete: lacking attribute values, lacking certain attributes of interest, or containing onlyaggregate data

   Noisy: containing errors or outliers

   Inconsistent: containing discrepancies in codes or names

2. Tasks in data preprocessing

   Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolveinconsistencies.

   Data integration: using multiple databases, data cubes, or files.

   Data transformation: normalization and aggregation.

   Data reduction: reducing the volume but producing the same or similar analytical results.

   Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Data Cleaning

Data cleaning is a technique deal with detecting and removing inconsistencies and error from the data in-order to get better quality data. Data cleaning is performed as a data preprocessing step

while preparing the data for a data warehouse. Good quality data requires passing a set of quality criteria. Those criteria include: Accuracy, Integrity, Completeness, Validity, Consistency, Uniformity, Density and Uniqueness.

Data Integration

Data Integration is a data preprocessing technique that takes data from one or more sources and mapping it, field by field onto a new data structure. Idea is to merge the data from multiple sources into a coherent data store. Data may be distributed over different databases or data warehouses. There may be necessity of enhancement of data with additional (external) data. Issues like entity identification problem.

Data Transformation

In data transformation data are consolidated into appropriate form to make suitable for mining, by performing summary or aggregation operations. Data transformation involves following:

- Data Smoothing
- Data aggregation
- Data Generalization
- Normalization
- Attribute Construction

Data Reduction

If the data set is quite huge then the task of data mining and analysis can take much longer time, making the whole exercise of analysis useless and infeasible. Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

The data reduction strategies include:

- Data cube aggregation
- Dimensionality reduction
- Data discretization and concept hierarchy generation
- Attribute Subset Selection

Similarity measures

The similarity measure is the measure of how much alike two data objects are. Similarity measure in a data mining context is a distance with dimensions representing features of the objects. If this distance is small, it will be the high degree of similarity where large distance will be the low degree of similarity.

The similarity is subjective and is highly dependent on the domain and application. For example, two fruits are similar because of color or size or taste. Care should be taken when calculating distance across dimensions/features that are unrelated. The relative values of each element must

be normalized, or one feature could end up dominating the distance calculation. Similarity are measured in the range 0 to 1 [0,1].

Two main considerations about similarity:

- Similarity = 1 if X = Y (Where X, Y are two objects)
- Similarity = 0 if X ≠ Y

Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering. As the names suggest, a similarity measures howclose two distributions are.

Similarity Measure

Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1(complete similarity)

Dissimilarity Measure

Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different)

Proximity

Refers to a similarity or dissimilarity.

Summary statistics

Summary statistics summarize and provide information about your sample data. It tells you something about the values in your data set. This includes where the average lies and whether your data is skewed. Summary statistics fall into three main categories:

- Measures of location (also called central tendency).
- Measures of spread.
- Graphs/charts

Summary Statistics: Measures of location

Measures of location tell you where your data is centered at, or where a trend lies. Following common measures of location for a full definition and examples for that particular measure:

- Mean (also called the arithmetic mean or average)
- Geometric mean (used for interest rates and other types of growth)
- Trimmed Mean (the mean with outliers excluded)
- Median (the middle of a data set)

Summary Statistics: Measures of spread

Measures of spread tell you (perhaps not surprisingly!) how spread out or varied your data set is. This can be important information. For example, test scores that are in the 60-90 range might be expected while scores in the 20-70 range might indicate a problem. Range isn't the only measure

of spread though. The names below for a particular measureof spread.

- Range (how spread out your data is?)
- Interquartile range (where the "middle fifty" percent of your data is?)
- Quartiles (boundaries for the lowest, middle and upper quarters of data)
- Skewed (does your data have mainly low, or mainly high values?)
- Kurtosis (a measure of how much data is in the tails?)

Summary Statistics: Graphs and Charts

There are literally dozens of ways to display summary data using graphs or charts. Some of the most common ones are listed below.

- Histogram
- Frequency Distribution Table
- Box plot
- Bar chart
- Scatter plot
- Pie chart

Data distributions

A **data distribution** is a function or a listing which shows all the possible values (or intervals) of the data. It also (and this is important) tells you how often each value occurs. Often, the data in a distribution will be ordered from smallest to largest, and graphs and charts allow you to easily see both the values and the frequency with which they appear.

From a distribution you can calculate the probability of any one particular observation in the sample space, or the likelihood that an observation will have a value which is less than (or greater than) a point of interest.

Data distributions are used often in statistics. They are graphical methods of organizing and displaying useful information. There are several types of data distributions like dot plots, histograms, box plots, and tally charts. Here we will focus on dot plots and histograms

Dot Plots

Dot plots show numerical values plotted on a scale. Each dot represents one value in the set of data. In the example below, the customer service ratings range from 0 to 9. The dots tell us the frequency, or rate of occurrence, of customers who gave each rating. If you look at the 5 rating, you can see that three customers gave that rating, and if you look at a score of 9, eight customers gave that rating. We can also see that ratings were provided by fifty customers, one dot for each customer.

Figure 1: Example of a dot plot

Now imagine that ratings were provided by five hundred customers. It would not be practical or useful to have a distribution of five hundred dots. For this reason, dot plots are used for data that have a relatively small number of values.

Histograms

Histograms display data in ranges, with each bar representing a range of numeric values. The height of the bar tells you the frequency of values that fall within that range. In the example below, the first bar represents black cherry trees that are between 60 and 65 feet in height. The bar goes up to three, so there are three trees that are between 60 and 65 feet.

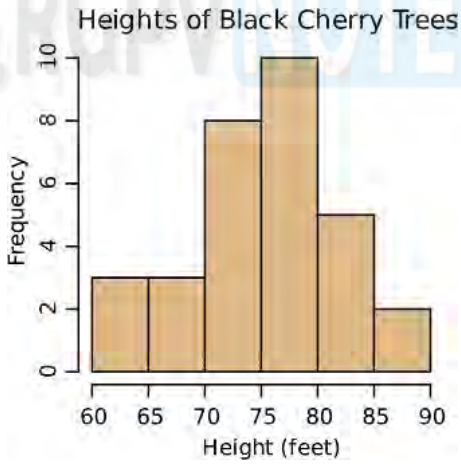

Figure 2: Example of a histogram

Histograms are an excellent way to display large amounts of data. If you have a set of data that includes thousands of values, you can simply adjust the frequency interval to accommodate a larger scale, rather than just 0-10.

Basic data mining tasks

The two "high-level" primary goals of data mining, in practice, are prediction and description.

1. **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
2. **Description** focuses on finding human-interpretable patterns describing the data.

The relative importance of prediction and description for particular data mining applications can vary considerably. However, in the context of KDD, description tends to be more important than prediction. This is in contrast to pattern recognition and machine learning applications (such as speech recognition) where prediction is often the primary goal of the KDD process.

The goals of prediction and description are achieved by using the following primary data mining tasks:

1. Classification is learning a function that maps (classifies) a data item into one of several predefined classes.
2. Regression is learning a function which maps a data item to a real-valued prediction variable.
3. Clustering is a common descriptive task where one seeks to identify a finite set ofcategories or clusters to describe the data.

   - Closely related to clustering is the task of probability density estimation which consists of techniques for estimating, from data, the jointmulti-variate probability density function of all the variables/fields in the database.

4. Summarization involves methods for finding a compact description for a subset of data.
5. Dependency Modeling consists of finding a model which describes significant dependencies between variables. Dependency models exist at two levels:

   1. The structural level of the model specifies (often graphically) which variables are locallydependent on each other, and
   2. The quantitative level of the model specifies the strengths of the dependencies usingsome numerical scale.

Change and Deviation Detection focuses on discovering the most significant changes in the data from previously measured or normative values.

Data Mining V/s Knowledge discovery in databases

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 3) shows data mining as a step in an iterative knowledge discovery

process.

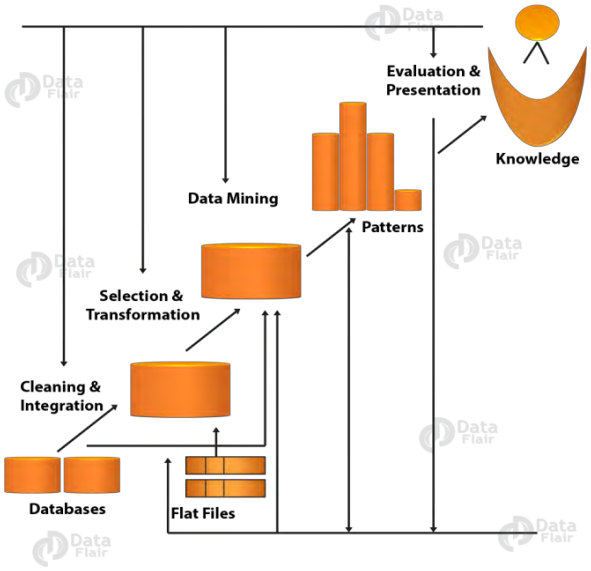

Figure 3: KDD process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- **Data cleaning**: Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
- **Data integration**: At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- **Data selection**: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- **Data transformation**: Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- **Data mining**: It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- **Pattern evaluation**: In this step, strictly interesting patterns representing knowledge are identified based on given measures.
- **Knowledge representation**: Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a

large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

Issues in Data mining

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.

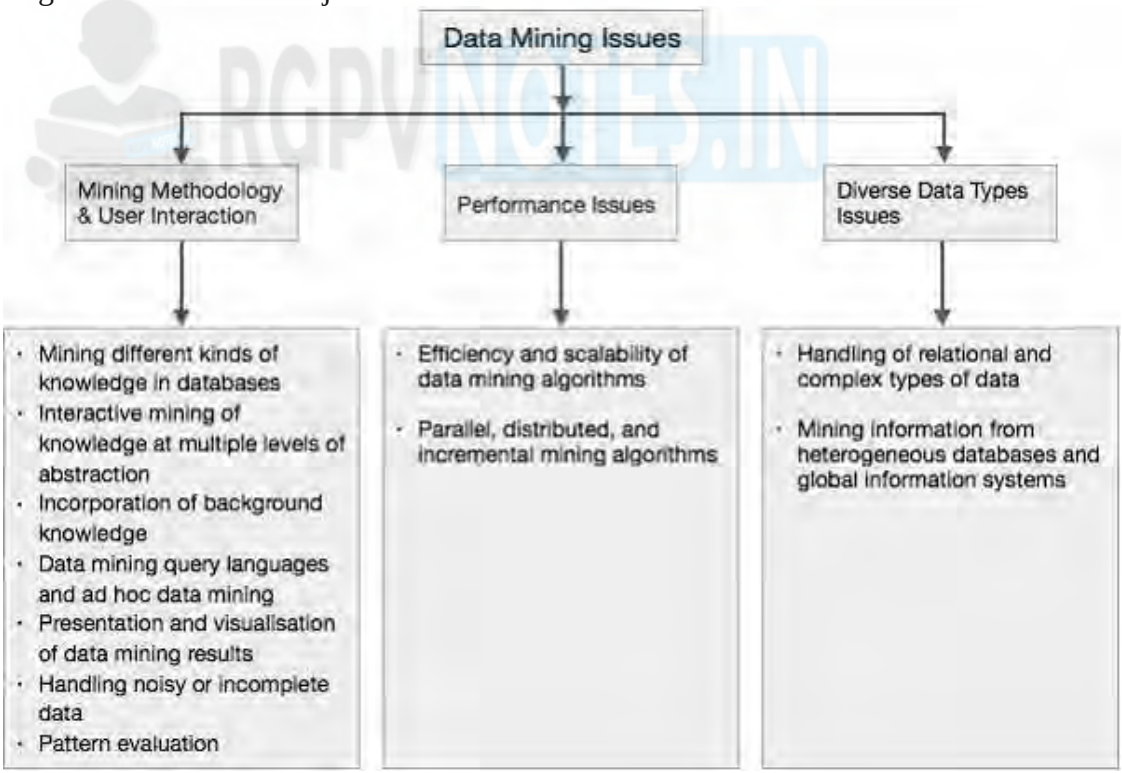

Figure 4: Issues in Data mining

Issues that need to be addressed by any serious data mining package are:

i. Uncertainty Handling

ii. Dealing with Missing Values

iii. Dealing with Noisy Data

iv. The Efficiency of Algorithms

v. Constraining Knowledge Discovered to only Useful

vi. Incorporating Domain Knowledge

vii. Size and Complexity of Data

viii. Data Selection

ix. Understandability of Discovered Knowledge: Consistency between Data and Discovered Knowledge

Data Mining System Classification

A data mining system can be classified according to the following criteria −

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines

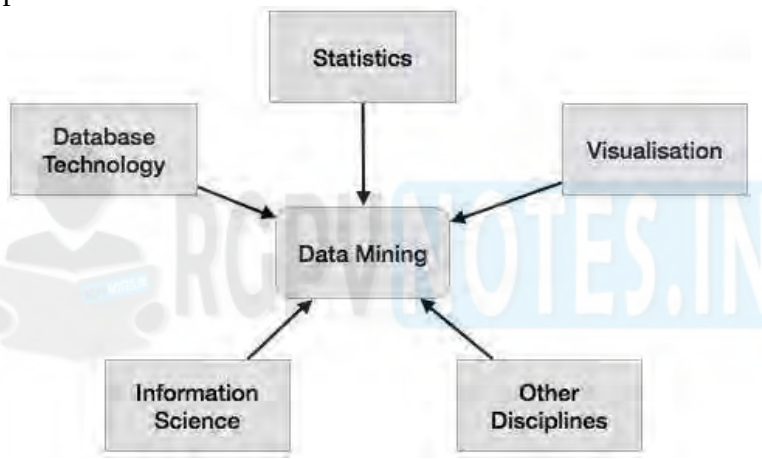

Figure 5: Data Mining System Classification

Data mining systems depend on databases to supply the raw input and this raises problems, such as the databases tend to be dynamic, incomplete, dynamic, noisy and large. Other problems arise as a result of the inadequacy and irrelevance of the information stored. The difficulties in data mining can be categorized as

a) Limited information
b) Noise or missing data
c) User interaction and prior knowledge
d) Uncertainty
e) Size, updates and irrelevant fields

Introduction to Fuzzy sets and Fuzzy logic:

The word fuzzy refers to things which are not clear or are vague. Any event, process, or function that is changing continuously cannot always be defined as either true or false, which means that we need to define such activities in a Fuzzy manner.

Fuzzy Logic resembles the human decision-making methodology. It deals with vague and imprecise information. This is gross oversimplification of the real-world problems and based on

degrees of truth rather than usual true/false or 1/0 like Boolean logic.

Take a look at the following diagram. It shows that in fuzzy systems, the values are indicated by a number in the range from 0 to 1. Here 1.0 represents absolute truth and 0.0 represents absolute falseness. The number which indicates the value in fuzzy systems is called the truth value.

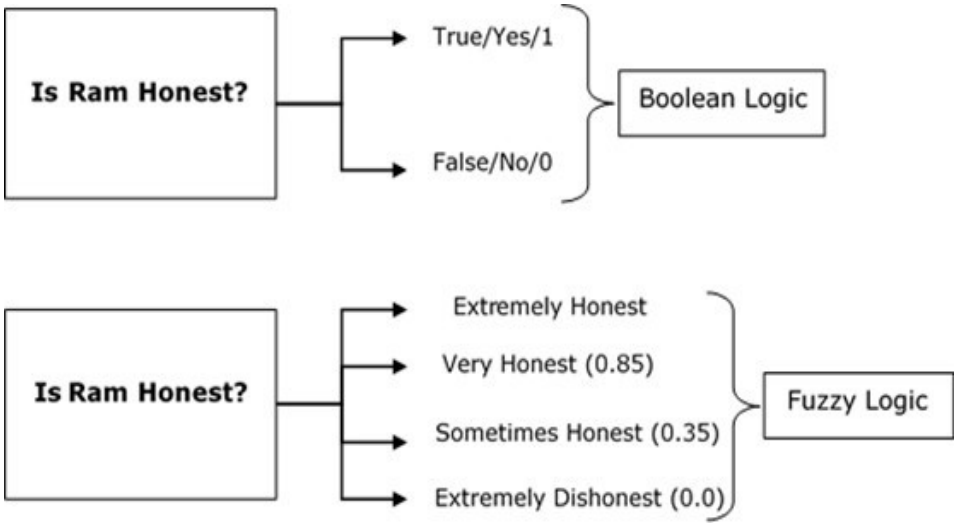

Figure 6: Fuzzy System

In other words, we can say that fuzzy logic is not logic that is fuzzy, but logic that is used to describe fuzziness. There can be numerous other examples like this with the help of which we can understand the concept of fuzzy logic.

A set is an unordered collection of different elements. It can be written explicitly by listing its elements using the set bracket. If the order of the elements is changed or any element of a set is repeated, it does not make any changes in the set.

Example:

A set of all positive integers.

A set of all the planets in the solar system.

A set of all the states in India.

A set of all the lowercase letters of the alphabet.

Mathematical Representation of a Set

Sets can be represented in two ways −

Roster or Tabular Form

In this form, a set is represented by listing all the elements comprising it. The elements are enclosed within braces and separated by commas.

Following are the examples of set in Roster or Tabular Form −

- Set of vowels in English alphabet, A = {a,e,i,o,u}
- Set of odd numbers less than 10, B = {1,3,5,7,9}

Set Builder Notation

In this form, the set is defined by specifying a property that elements of the set have in common. The set is described as A = {x:p(x)}

Example 1 − The set {a,e,i,o,u} is written as:

A = {x:x is a vowel in English alphabet}

Example 2 − The set {1,3,5,7,9} is written as:

B = {x:1 ≤ x < 10 and (x%2) ≠ 0}

If an element x is a member of any set S, it is denoted by x∈ S and if an element y is not a member of set S, it is denoted by y∉ S.

Example − If S = {1,1.2,1.7,2},1 ∈ S but 1.5 ∉ S

Cardinality of a Set

Cardinality of a set S, denoted by |S||S|, is the number of elements of the set. The number is also referred as the cardinal number. If a set has an infinite number of elements, its cardinality is ∞.

Example − |{1,4,3,5}| = 4,|{1,2,3,4,5,…}| = ∞

If there are two sets X and Y, |X| = |Y| denotes two sets X and Y having same cardinality. It occurs when the number of elements in X is exactly equal to the number of elements in Y. In this case, there exists a bijective function 'f' from X to Y.

|X| ≤ |Y| denotes that set X's cardinality is less than or equal to set Y's cardinality. It occurs when the number of elements in X is less than or equal to that of Y. Here, there exists an injective function 'f' from X to Y.

|X| < |Y| denotes that set X's cardinality is less than set Y's cardinality. It occurs when the number of elements in X is less than that of Y. Here, the function 'f' from X to Y is injective function but not bijective.

If |X| ≤ |Y| and |X| ≤ |Y| then |X| = |Y|. The sets X and Y are commonly referred as equivalent sets.

***

RGPVNOTES.IN

Thank you for using our services. Please support us so that we can improve further and help more people.
https://www.rgpvnotes.in/support-us

If you have questions or doubts, contact us on
WhatsApp at +91-8989595022 or by email at hey@rgpvnotes.in.

For frequent updates, you can follow us on
Instagram: https://www.instagram.com/rgpvnotes.in/.