**What is data mining and what are different sources of data in modern world?**

**Data Mining:**

As mining in Data Mining suggests extracting something valuable. Data Mining is a process of discovering patterns in larger data sets. it is a process of finding anomalies, patterns, and correlations within large datasets to predict outcomes. In short, we can say, data mining is a process used to extract usable data from larger sets of raw data. Following are some sources of data in the modern world:

Web Data

E-Commerce

Bank Transactions

Digital Media

Online Games

Research and Science Biomedical

Social media data

Youtube

Cloud

Officials Statistics

Weather Forecast

Public datasets for Machine Learning etc.

**Q2: How data mining is different from DBMS?**

Because in Data Mining we need implicit knowledge. But in DMBS we don't need any implicit knowledge get extract or get data. AS we Know, DBMS is an organized collection of data where we write a query for required data and get the information that we want which is called raw data. On the other hand, Data Mining analyzes data from different information to discover usable information and previously unknown information from raw data.

**Q3: How can we extract knowledge from data?**

**Steps to extract knowledge from raw data**

There are seven Steps to Extract Knowledge from raw data:

Defining the problem

Building the database

Use of data

Preparation

Building the Model

Assessment Model

Interpretation of Model results and improvement

**Q4: What are different types of data and which tasks can we perform on data?**

**Different Types Of Data:**

Time Related or Sequence Data

Data Streams

Spatial Data

Engineering Design

HyperText and Multimedia Data

Web Data

**The task to be performed on data:**

Characterization

Discrimination

Association

Clustering

Trend

Deviation and outer Analysis

**Q5: State applications of data mining in real world**

Applications of Data Mining in Real World:

Data mining is used in almost every field of life. Here are some field where it used:

Retail

Telecommunication

Banking

Fraud Analysis

DNA mining

Stock market Analysis

Web Mining

Weblog Analysis

Lie Detection

Customer Segmentation

Financial banking

Corporate Surveillance

Research Analysis

Criminal investigation

Bioinformatics

Service provider

**Q6: What kind of data can be mined?**

Following are some types of data that can be mined:

1. Flat Files:

Flat files are defined as data files in text form or in binary form with a structure that can be extracted by data mining algorithms.

2. Relational Database:

A relational database is a collection of data in an organized way.

3. DataWarehouse:

A data warehouse is a database system designed for analytics. Data warehousing is the process of extracting and storing data that allow easier reporting. Data mining is generally considered as the process of extracting useful data from a large set of data.

4. Transactional Databases:

Transactional database is a collection of data organized by date, timestamps, etc. to represent transactions in databases.

5. Multimedia Databases:

This is the type of database that is used to store multimedia data like videos, audio, images, etc.

6. Spatial Databases:

Spatial data is associated with geographic locations such as Countries, States, cities, towns, etc.

7. Time Series Databases:

A time series is a sequence of data points recorded at particular time points - most often in regular time intervals (like years, months, weeks, days, min, etc.).

8. World Wide Web(WWW):

Web mining can define as the method of utilizing data mining techniques and algorithms to extract useful information directly from the web, such as Web documents and services, hyperlinks, Web content, and server logs.

**Q7: What is data warehouse and what operations can we perform on data in data warehousing?**

Data Warehouse:

Repository of information collected from multiple sources, stored under a unified schema, usually resides at a single point. Or we can say, The data warehouse is referred to as a place where meaningful and useful data can be stored. Data from different organizations or sources are added to the warehouse so it can be fetched and conformed for delete errors.

Following are some operations which are performed on data in data warehousing:

Data Cleaning

Data Transforming

Data integration

data Loading

Periodic Refreshing

Q8: Explain different patterns of descriptive data?

Different Patterns of Descriptive data:

There are three major categories of Descriptive data:

Clustering

Association

Summarization

1. Clustering:

The process in which we group objects without knowing their class label. In this process, we are maximizing intra-class similarities of objects while minimizing inter-class similarities of objects.

2. Association:

The process in which we are going to predict the associated object of the previous one As it is a prediction of the most associated objects of the current object.

3. Summarization:

Describes a given set of data in a concise and summative manner, presenting general interesting properties Following terms are most likely to associate this pattern of descriptive data

Data Characterization: To characterize data as useful or not useful Data Discrimination: Differentiate data from the one which is not valuable.

**Q9: Which patterns are to be minded in predictive data?**

Patterns to be mined in Predictive data:

Followings are the patterns to be mined in predictive data:

Classification

Predictions

Time-series Analysis

regression

Outlier Analysis

**Q10: Which technologies are used in data mining?**

Following are some technologies that are used in Data mining: ..

1. Statistics:

In Statistics, we deal with the collection, organization, computing of data, and its representation.

2. Information retrieval:

It deals with the retrieval of information from different sources like documents etc.

3. Machine Learning:

In ML, we deal with the performance of a computer, how computers can be efficient on the basis of data.

4. Data warehousing:

It deals with the collection of data from different databases for analytics.

5. Database Systems:

As a data warehouse collects data from different databases so it means a lot in data mining.

6. High-performance computing:

These are responsible for storing a large amount of mined data accurately and fastly.

7. Algorithms:

Algo's deals with the working of different operations as algorithms have lots of importance in computer science.

8. Pattern recognition:

it is used to make patterns instead of searching each and every data from the database.

9. Visualization:

it deals with output data to show it in a user-friendly mode.

**Q11: State different applications of data mining?**

Applications of data mining:

There are many fields where Data Mining comes into play But we discuss here two fields.

1. Business Intelligence:

Business Intelligence is the technology that deals with the analytic and predicted operations of the business. Without data mining, effective marketing is not possible. .. Some common examples that come under this field Application of Data Mining:

Stock Market

Basket product prediction

Choice of Business which gives more profit

Prediction and analysis of sales

2. Search Engines:

The place (servers) where we can search out our desired data or information. It may include the internet such as google search engines where we can search for different types of things of our choice and we can get a suggestion out of which we searched this is all due to data mining. Some common examples that come under this field are as follow:

WWW

Web Analysis

Hyperlink Analysis

12: Discuss major challenges and issues in data mining?

Following are the major challenges and issues in data mining:

Mining Methodology

User Interaction

Efficiency

ScaleAbility

Diversity of database types

Data mining and society

1. Mining methodology:

it may include the following points:

Mining Various kind of data

Multidimensional space data

Handle noise, uncertainties, and incomplete data.

2. User interaction:

it may include the following points:

Background knowledge needs to incorporate.

Present and visualization of data.

Interactive mining

In General, The application developed using data mining must be user-friendly it must be suitable for users to work with without knowing the implicit details of data mining.

3. Efficiency/Scalability:

Parallel,Distrubuted,stream and intreamental mining methods.

The application developed using data mining should be Efficient

4.Data Diversity:

it may include the following points:

Dealings with complex types of data

Mining dynamic, networked, and global data repository

As the collected data is in many data types so in order to apply data mining operation on data it must be in the same data type so it's really challenging.

5. Impact on Society:

it may include the following points:

Social impacts of data mining

privacy-preserving data minng

invisible data mining

Q13: What are the data objects considered while preparing data?

Data sets are made up of data objects. A data object represents an entity. Data objects are basically storage regions where a set of relative information is stored.

Following are some data objects considered while preparation of data:

1. Records:

Collection of fields of different data types, in fixed number and sequence.

2. Network and Graphs: Network:

A set of nodes with edges example WWW or other web pages.

3. Ordered data:

Categorical, Statistical data type where the variables have natural, ordered categories with specific distances between them. It's an ordered data with time series for example Video etc.

4. Spatial data and images:

It's geographical data an example maps, images, videos, and their sequences, etc.

Q14: Define structured data and its constituents?

Structured data:

Structured data is data that has clear, definable relationships between the data points, with a pre-defined model containing it. It is the data that is both highly organized and easy to digest making analytics possible through the use of legacy data mining.

Constituents:

Following are some constituents of Structured Data:

1. Dimensionality:

It is the number of rows and columns.

2. Sparsity:

It is no of cells that are empty means the cells having no data.

3. Dimensions/Density:

It is opposite to the sparsity of cells having data inside them.

4. Resolutions:

It is the pattern that is responsible for making dimensional data resolved.

5. Distribution:

List of functions showing all the possible values of the data and how often they occur

Q15: What can be the architecture or storage mechanism of data when it is prepared for data mining task?

The architecture of data:

when it is prepared for the data mining tasks consists of the following components:

1. Data objects:

It is a storage region where a set of values or attributes are stored.it represents an entity. they are also known as with the following names like tuples, samples, data points, and instances, etc. Some examples of data objects are the Sales database, Medical Database, and University Database, etc.

2. Data Matrix:

It stores data in the form of MxN dimensions.Where M represents rows while N represents columns. Rows of Data matrix represents Objects while a column of Data matrix represents Attributes.

Q16: What are attributes?

Attributes:

A data field represent a characteristic or feature of the data object.

other names: Dimension, Features, Variables

Q17: Explain different types of attributes?

Types of Attributes:

Nominal

Binary

Numeric

1. Nominal:

Common Data Type, These types of attributes hold data value that is categorical i.e. Assitant,sub-Assistant, etc.

2. Binary:

Special case of Nominal. These types of attributes have a data value only has two options i.e. Male and Female.

3. Numeric:

These type of attributes have a data value of numbers data type i.e ordinal scale or ratio scale.

Q18: Differentiate between discrete and continuous attribute?

Differentiate between discrete and continuous attribute:

Discrete Attributes:

The attributes that have only a finite or countably infinite set of values. In General, those samples are countable. For example, Binary value, no of words in a document, professions and no of pages in books, etc.

Continuous Attributes:

The attributes that have real numbers as attribute values or samples that can take any value including decimal points i.e. temperature or height etc.

Q19: What is data visualization? Discuss its advantages?

Data Visualization:

Transforms data into images for an effective and accurate presentation or for better understanding and Analysis.

Advantages of data visualization:

Questions Answers

Make a decision

See data in the context

Tell a story

Analysis and discover

Present an argument

Inspire

Q20: What are different data visualization methods?

Data Visualization Methods:

Following are the methods of data visualization

1. Pixel Oriented:

Represent many objects on the same screen at the same time. Map each data value to a pixel of the screen and arrange pixels adequately. The smallest unit of the image equals the size of the smallest cell in the matrix form. It helps to organize images well.

2. Geometric Visualization:

Visualization of geometric transformations and projection of the data.

3. Icon Visualization:

It represents the data features in the form of icons.

4. Hierarchical Visualization:

It is the representation of data in the form of hierarchy like in the tree form

5. Complex data visualization:

It is the representation of hyperlink data like data of social media data.

Q21: What is pixel based data visualization approach?

Pixel-based data visualization is basically the representation of all the objects on a single screen. Objects are uniformly arranged on the screen. Pixels are uniformly arranged on the screen and show data. Different values are shown in pixel colors.

Q22: What are statistical measures of data?

Statistical Measures:

The procedure of understanding, interpreting, and summarizing of data.

Following are some measuring characteristics of Statistical measures:

1. Central Tendency:

Central value of data( mean value of data )

2. Variation:

How data deviate from the mean point. Or How data values differ from each other.

3. Spread:

How much data is spread from a to a point.

Q23: What is central tendency?

Central Tendency:

A single score that defines the middle/Center of the distribution. In Statistics, A central tendency is a center or typical value for a probability distribution.

Following are the best representation of Central tendency:

Mean

Median

Mode

Q24: Differentiate between symmetric and skewed data?

Differentiate between symmetric and skewed data:

Symmetric data:

Symmetric data that has an equal no of data points on either side or equally distributed.

Skewed or Asymmetric data:

Asymmetric data is data that is differ from a normal or gaussian distribution.This may be positively skewed data or negatively skewed data.

Positively Skewed data: Having concentrated data on the right side of the middle data point.

Negatively Skewed Data: Having concentrated data on the left side of the middle point of distribution.

Q25: What is dispersion of data?

Dispersion of data:

The measure of the extent to which individual items vary. also called how much data is spread or scattered. It indicates the consistency of data. When the data set has a large value it has large scattered values and when the data set has a small set of values it has small scattering values. It is used to measure the variability of data used in data mining.

Q26: Define Box Plot Analytics?

Box plot:

Boxplots are the visual representation of quartiles. If we want to see the graphical forms of data quartiles it would be in the form of Boxplots. Boxplot is a standard way of displaying data based on a five-number summary. The end of the box is at the 1st and 3rd quartile and the median is within the box.

Q27: What are Five Number Summary of Box Plot?

Five Number Summary of Box Plot:

A boxplot displays the five-number summary of a set of data. The five-number summary is:

Minimum

1st quartile (Q1)

Median

3rd quartile (Q3)

Maximum

Q28: What are Whiskers and Out liners?

Whiskers:

Two lines outside the box extended to Min & Max.

Outliers:

Points beyond a specified outlier threshold plotted individually or Outliers may be plotted as individual points.

Q29: what are histograms?

Histograms:

The histogram is a measure of the graphical display of tabulated values/frequencies. No of the data items in intervals are Frequencies. And bar height describes the occurrence of the data items.

Q30: Difference between Bar Chart & Histogram?

Difference between Bar Chart & Histogram

Bar Chart:

Barcharts may have gaps mean they may not be continuous.

Barcharts are used for categorical data.

Histogram:

Histograms are always continued and defined for every data point.

Histograms are used for nominal data.

Q31: Types of Histogram?

Types of Histogram:

The following are the types of Histogram:

1. Bell Shaped:

The Normal Distribution. In this type, the points on one side of the middle are as likely to occur as on the other side of the middle.

2. Doubled Peaked:

Suggests two Distributions Also Known as a bimodal distribution. It has two peaks. In this type, the data should be separated and analyzed as separate normal distributions.

3. Skewed:

Look for Other Processes in the Tail. It may be left-skewed or right-skewed.

4. Truncated:

Looks For reasons for the sharp end of distribution or pattern.

5. Ragged Plateau:

No Single clear process or pattern. Also known as random distribution. It lacks an apparent pattern and has several peaks. It can be the case that different data properties were combined. Therefore, the data should be separated and analyzed separately.

Q32: Define Quantile- Quantile Plots?

Quantile- Quantile Plots:

The quantile-quantile (Q-Q) plot is a graphical technique for evaluating if two sets of data come from widely disseminated populations. Or A quantile-quantile plot is a probability plot in statistics and is a graphical way to compare two probability distributions by plotting their quantities against each other.

Q33: what is scatter Plot?

Scatter Plot

A scatter plot is a type of plot which is used to display values for typically two variables for a set of data. E.g Graph between height and weight.

Correlation: It means How two data values are related to each other. The variables may be Positively (directly) related or Negatively (inversely) related.

Clusters of points: We check data where two points having clusters.

Q34: what is geometric visualization?

Geometric visualization:

In Geometric visualization, we focus on the iransformation of data and projection of data. It means that how data is transformed from one domain to another domain and what is the projection of that data.

Q35: what is icon-based visualization?

Icon-based visualization:

Icon Based visualization uses small icons to represents multi-dimension data values.

Typical visualization methods

Chernoff Faces

Stick Figures

General Techniques:

1. Shape coding:

This technique, uses shapes to represent certain information encoding

2. Color icons:

This technique, uses color icons to encode more information

3. Tile bars:

This technique, uses small icons to represent the relevant feature vectors in document retrieval

Q36: what is Chernoff faces?

Chernoff faces:

Represent variables in 2-dimension space. Chernoff faces, invented in 1973 by Herman Chernoff, show the form of a human face with multivariate details. The individual pieces, such as the eyes, ears, mouth, and nose, by their form, size, location, and orientation, represent the values of the variables. The concept behind using faces is that, without trouble, people quickly identify faces and note slight changes. The Chernoff faces treat each variable differently. Since the characteristics of the faces differ in perceived significance, it is important to carefully select the way in which variables are mapped to the characteristics.

Q37: what is Hierarchical Visualization?

Hierarchical Visualization:

It is the representation of data in the form of hierarchy like in the tree form. If you are looking to view clusters of information, particularly if they flow from a single origin point, hierarchical visualizations are best suited. The downside to these diagrams is that they appear to be more complicated and hard to understand, which is why the tree diagram is most commonly used. It is the simplest to follow due to its linear path.

Examples of hierarchical data visualizations include:

Tree diagrams

Ring charts

Sunburst diagrams

Q38: what is Dimensional Stacking?

Dimensional Stacking:

In Dimensional Staking the word Stack or Stacking means to place one thing to another thing. So we can easily say that Dimensional Stacking is about placing dimension to another dimension.

Proper Definition:

Dimensional Stacking is a technique for displaying multivariate data in two-dimensional screen space. This technique involves the discretization and recursive embedding of dimensions, each resulting N-dimensional bin occupying a unique position on the screen.

Important Attributes are kept on an outer level.

Q39: what is Tree Mapping in hierarchical visualization?

Tree Mapping:

In Tree mapping, we generate data in the form of rectangles. Its simple idea is just like a tree in which we do screen Filling like we have a rectangular box and we keep placing small rectangles in it. The small rectangle means that we will see the root node then their child node and so on.

Proper Definition:

Treemaps are visualizations for hierarchical data. They are made of a series of nested rectangles of sizes proportional to the corresponding data value. A large rectangle represents a branch of a data tree, and it is subdivided into smaller rectangles that represent the size of each node within that branch.

Q40: what is N-vision and Auto-Visual?

N-vision:

N-vision is company of data collection that collects data for computer vision or image processing. Dynamic interaction through data gloves and stereo Display.

Auto-Visual:

Auto visual is a static interaction of users with data in the form of queries.