Q.1 Define data mining. Classify data mining systems.

Ans. Data mining → Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.

→ There are classification of data mining systems :-

- Classification based on the mined databases
- Classification based on the type of mined knowledge.
- Classification based on statistics
- Classification based on Machine Learning
- Classification based on Visualization
- Classification based on information science.
- Classification based on utilized techniques.
- Classification based on adapted applications.

→ Classification Based on the mined databases → A data mining systems can be classified based on the type of databases system can be further segmented based on distinct principles, such as data models, types of data, etc. which further assist in classifying a data mining system.

→ Classification based on the type of knowledge Mined → A data mining system categorized based on the kind of knowledge mined may have the following functionalities :-

1. Characterization
2. Association and correlation Analysis
3. Classification
4. Prediction
5. Outlier Analysis
6. Evolution Analysis

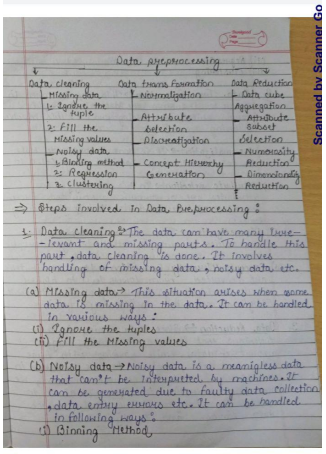
→ Classification Based on the Techniques Utilized → A data mining system can also be classified based on the type of techniques that are being incorporated. These techniques can be assessed based on the involvement of user interaction involved or the methods of analysis employed.

→ Classification Based on the Applications Adapted → Data mining systems classified based on adapted applications adapted are as follows :-

1. Finance
2. Telecommunications
3. DNA
4. Stock Markets
5. E-mail

Q.2 Explain data preprocessing steps.

Ans Preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



→ Steps Involved in 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 :-
(i) Ignore the tuples
(ii) Fill the missing values

(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 :-
(i) Binning Method

(ii) Regression
(iii) Clustering

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

(a) 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)
(b) Attribute selection → In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
(c) Discretization → This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
(d) Concept Hierarchy Generation → Here attributes are converted from lower level to higher level in hierarchy. For example - The attribute "city" can converted to "country".

3. Data Reduction :- Since data mining is a technique that is used to handle huge amount of data, while working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are :-

(a) Data cube aggregation → Aggregation operation is applied to data for the construction of the data cube.
(b) Attribute subset selection → The highly relevant attribute should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p-value of the attribute. The attribute having p-value greater than significance level can be discarded.
(c) Numerosity Reduction → This enable to store the model of data instead of whole data, for example :- Regression Models.
(d) Dimensionality Reduction → This reduce the size of data by encoding mechanisms. It can be lossy or lossless.

Q.3 What are the major issues in data mining? Write advantage of data mining system.

Ans The major issues regarding :-
→ Mining Methodology and User interaction
→ Performance issues
→ Diverse Data types issues

1. → Mining Methodology and user interaction issues It refers to the following kinds of issues-

- Mining different kinds of knowledge in databases - Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- Interactive mining of knowledge at multiple levels of abstraction → The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- Incorporation of background knowledge → To guide discovery process and to express the discovered patterns, the background knowledge can be used.
- Data mining query languages and ad hoc data mining → Data mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- Presentation and visualization of data mining results → Once the patterns are discovered it needs to be expressed in high level languages, and visual representations.
- Handling noisy or incomplete data → The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities.
- Pattern evaluation → The pattern discovered should be interesting because either they represent common knowledge or lack novelty.

2. → Performance issues :- There can be performance related issues such as follows-

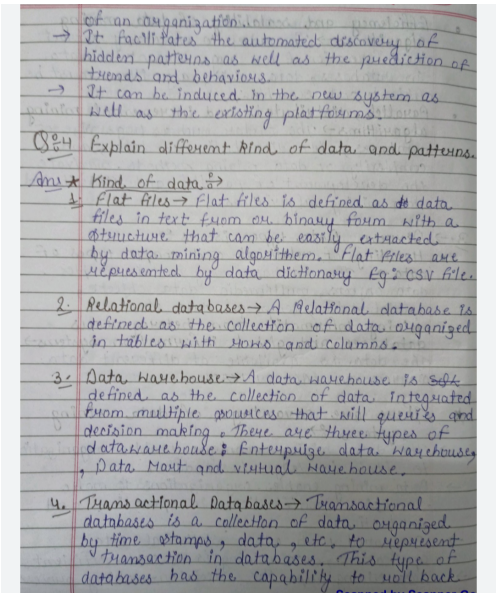
- Efficiency and scalability of data mining algorithms → In order to effectively extract the information from huge amount of data in databases, data mining algorithms must be efficient and scalable.
- Parallel, distributed and incremental mining algorithms → The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms.

3. → Diverse Data types issues :-

- Handling of relational and complex types of data → The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc.
- Mining information from heterogeneous databases and global information systems → The data is available at different data sources on LAN or WAN.

★ There are advantage of data mining systems :-

→ The data mining technique enables organization 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 behaviours.
→ It can be induced in the new system as well as the existing platforms.

Q4) Explain different kind of data and patterns.

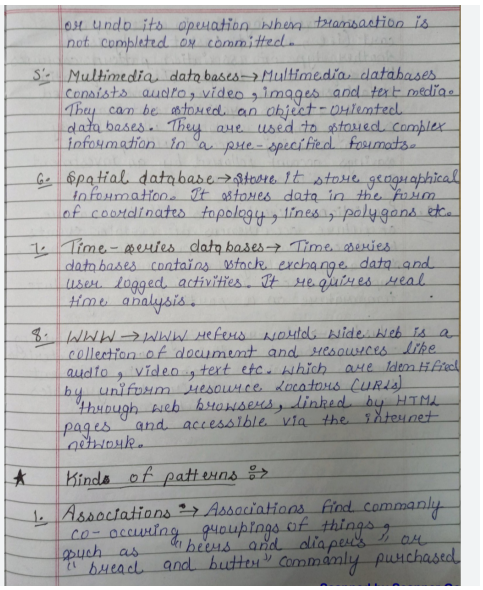
Ans ★ Kind of data :-

1. Flat files → Flat files is defined as the data files in text form or binary form with a structure that can be easily extracted by data mining algorithms. Flat files are represented by data dictionary Eg: CSV file.

2. Relational databases → A Relational database is defined as the collection of data organized in tables with rows and columns.

3. Data Warehouse → A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making. There are three types of data warehouse: Enterprise data warehouse, Data Mart and virtual warehouse.

4. Transactional Databases → Transactional databases is a collection of data organized by time stamps, data, etc. to represent transaction in databases. This type of databases has the capability to roll back


or undo its operation when transaction is not completed or committed.

5. Multimedia data bases → Multimedia databases consists audio, video, images and text media. They can be stored on object-Oriented databases. They are used to stored complex information in a pre-specified formats.

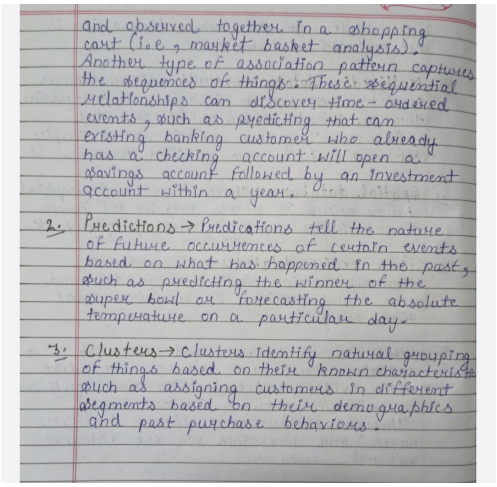
6. Spatial database → Store. It store geographical information. It stores data in the form of coordinates topology, lines, polygons etc.

7. Time-series data bases → Time series databases contains stock exchange data and user logged activities. It requires real time analysis.

8. WWW → WWW refers world wide web is a collection of document and resources like audio, video, text etc. which are identified by uniform resource locators (URIs) through web browsers, linked by HTML pages and accessible via the Internet network.

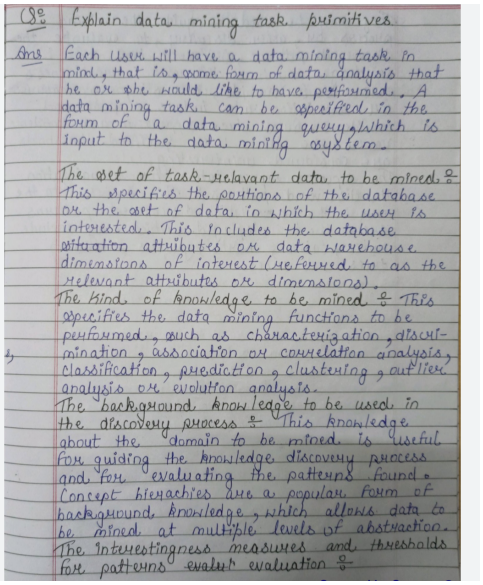
★ Kinds of patterns :-

1. Associations :- → Associations find commonly co-occuring groupings of things, such as "beers and diapers" or "bread and butter" commonly purchased


and observed together in a shopping cart (i.e, market basket analysis). Another type of association pattern captures the sequences of things. These sequential relationships can discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

2. Predictions → Predications tell the nature of future occurrences of certain events based on what has happened in the past, such as predicting the winner of the super bowl or forecasting the absolute temperature on a particular day.

3. Clusters → Clusters identify natural grouping of things based on their known characteristics such as assigning customers in different segments based on their demographics and past purchase behaviours.


Q5) Explain data mining task primitives.

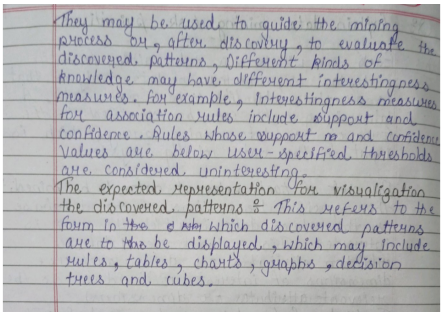
Ans Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed. A data mining task can be specified in the form of a data mining query, which is input to the data mining system.

The set of task-relevant data to be mined :- This specifies the portions of the database or the set of data in which the user is interested. This includes the database attribute attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

The kind of knowledge to be mined :- This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis or evolution analysis.

The background knowledge to be used in the discovery process :- This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierachies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction.

The interestingness measures and thresholds for patterns evalu evaluation :-


They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

The expected representation for visualization the discovered patterns :- This refers to the form in the ~~d~~ which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees and cubes.

##Data sets and Attributes

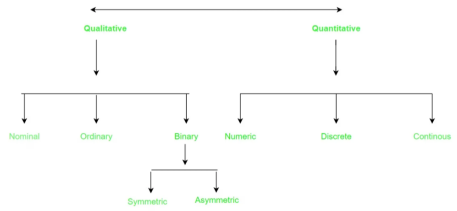
Data objects are the essential part of a database. A data object represents the entity. Data Objects are like a group of attributes of an entity. For example, a sales data object may represent customers, sales, or purchases. When a data object is listed in a database they are called data tuples.

Attribute:
It can be seen as a data field that represents the characteristics or features of a data object. For a customer, object attributes can be customer Id, address, etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector.

Type of attributes :

This is the First step of Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is the description of attribute types.

1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
2. Quantitative (Numeric, Discrete, Continuous)



Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as categorical attributes and there is no order (rank, position) among values of the nominal attribute.
Example :

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

2. Binary Attributes: Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.

Symmetric: Both values are equally important (Gender).
Asymmetric: Both values are not equally important (Result).

Attribute	Values
Gender	Male , Female

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

3. Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, interval, and ratio.

- An interval-scaled attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points.
- A ratio-scaled attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value.

2. Discrete : Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example :

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 .....etc

#What is Data Visualization?

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

 # Integration of a Data Mining System with a Database or Data Warehouse System :-

The data mining system is integrated with a database or data warehouse system so that it can do its tasks in an effective presence. A data mining system operates in an environment that needed it to communicate with other data systems like a database system. There are the possible integration schemes that can integrate these systems which are as follows −

No coupling − No coupling defines that a data mining system will not use any function of a database or data warehouse system. It can retrieve data from a specific source (including a file system), process data using some data mining algorithms, and therefore save the mining results in a different file.

Such a system, though simple, deteriorates from various limitations. First, a Database system offers a big deal of flexibility and adaptability at storing, organizing, accessing, and processing data. Without using a Database/Data warehouse system, a Data mining system can allocate a large amount of time finding, collecting, cleaning, and changing data.

Loose Coupling − In this data mining system uses some services of a database or data warehouse system. The data is fetched from a data repository handled by these systems. Data mining approaches are used to process the data and then the processed data is saved either in a file or in a designated area in a database or data warehouse. Loose coupling is better than no coupling as it can fetch some area of data stored in databases by using query processing or various system facilities.

Semitight Coupling − In this adequate execution of a few essential data mining primitives can be supported in the database/datawarehouse system. These primitives can contain sorting, indexing, aggregation, histogram analysis, multi-way join, and pre-computation of some important statistical measures, including sum, count, max, min, standard deviation, etc.

Tight coupling − Tight coupling defines that a data mining system is smoothly integrated into the database/data warehouse system. The data mining subsystem is considered as one functional element of an information system.

Data mining queries and functions are developed and established on mining query analysis, data structures, indexing schemes, and query processing methods of database/data warehouse systems. It is hugely desirable because it supports the effective implementation of data mining functions, high system performance, and an integrated data processing environment.