Important Instructions to examiners:

1) The answers should be examined by key words and not as word-to-word as given in the model answer scheme.
2) The model answer and the answer written by candidate may vary but the examiner may try to assess the understanding level of the candidate.
3) The language errors such as grammatical, spelling errors should not be given more Importance (Not applicable for subject English and Communication Skills.
4) While assessing figures, examiner may give credit for principal components indicated in the figure. The figures drawn by candidate and model answer may vary. The examiner may give credit for any equivalent figure drawn.
5) Credits may be given step wise for numerical problems. In some cases, the assumed constant values may vary and there may be some difference in the candidate's answers and model answer.
6) In case of some questions credit may be given by judgement on part of examiner of relevant answer based on candidate's understanding.
7) For programming language papers, credit may be given to any other program based on equivalent concept.

Q. No.	Sub Q. N.	Answer	Marking Scheme
1.	**a)**	**Attempt any three of the following:**	**(3×4=12)**
	a)	**Define. 1) Data mining** **2) Data warehousing.**	**4M**
	Ans:	**1. Data Mining**: Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web. **2. Data warehousing**: Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.	***(Data Mining: 2 marks, Data Warehousing: 2 marks)***
	b)	**Describe data cleaning techniques.**	**4M**
	Ans:	Data cleaning is performed as data preprocessing step while preparing the data for a data warehouse. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Noise is a random error or variance in a measured variable. Given a numerical attribute such as, say, price, we can "smooth" out the data to remove the noise by the following data smoothing techniques: **1. Binning:** Binning methods smooth a sorted data value by consulting its "neighborhood," that is, the values around it. The sorted values are distributed into a	***(List: 1 mark, Explanation of any 2 techniques: 1 ½ marks each)***

number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. Figure 3.4 illustrates some binning techniques. In this example, the data for price are first sorted and then partitioned into equal-frequency bins of size 3 (i.e., each bin contains three values). In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value. In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be equal-width, where the interval range of values in each bin is constant.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

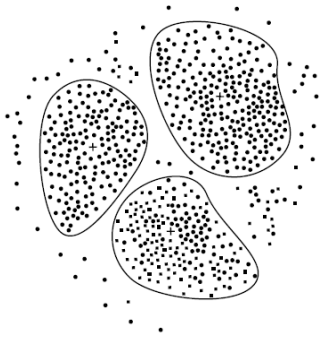
Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Figure: Binning

2. Regression: Data can be smoothed by fitting the data to a function, such as with regression. Linear regression involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers.



	c)	**Describe concept hierarchies.**	**4M**
	Ans:	A concept hierarchy for a given numerical attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior). Although detail is lost by such data generalization, the generalized data may be more meaningful and easier to interpret. This contributes to a consistent representation of data mining results among multiple mining tasks, which is a common requirement. In addition, mining on a reduced data set require fewer input/output operations and is more efficient than mining on a larger, un-generalized data set. Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining as a preprocessing step, rather than during mining.Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis. Five methods for concept hierarchy generation are defined below- binning histogram analysis entropy-based discretization and data segmentation by ―natural partitioning Binning: Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies. **Histogram Analysis:** Histograms can also be used for discretization. Partitioning rules can be applied to define range of values. The histogram analyses algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre specified number of concept levels have been reached. A minimum interval size can be used per level to control the recursive procedure. This specifies the minimum width of the partition, or the minimum member of partitions at each level. **Cluster Analysis:** A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a **node** of a concept hierarchy, where all noses are at the same conceptual level. Each cluster may be further decomposed into sub-clusters, forming a lower level in the hierarchy. Clusters may also be grouped together to form a higher-level concept hierarchy. **Segmentation by natural partitioning:** Breaking up annual salaries in the range of into ranges like ($50,000-$100,000) are often more desirable than ranges like ($51, 263, 89-$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform ―natural intervals. In general the rule partitions a give range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit.The rule can be recursively applied to each interval creating a concept hierarchy for the given numeric attribute.	***(Explanation : 4 marks)***
	d)	**Describe concept description in data mining.**	**4M**
	Ans:	**Concept Description:** Concept description refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways 1. **Data Characterization –** this refers to summarizing data of class under study. This class under study is called as Target Class.	***(Explanation : 4 marks)***

		2. Data Discrimination – It refers to the mapping or classification of a class with some predefined group or class. The simplest kind of descriptive data mining is concept description. A concept usually refers to a collection of data such as frequent buyers, graduate students, and so on. As a data mining task, concept description is not a simple enumeration of the data. Instead, concept description generates descriptions for characterization and comparison of the data. It is sometimes called class description, when the concept to be described refers to a class of objects. Characterization provides a concise and succinct summarization of the given collection of the data, while concept or class comparison (also known as discrimination) provides discriminations comparing two or more collections of data. Since concept description involves both characterization and comparison, techniques for accomplishing each of these tasks will study. Concept description has close ties with the data generalization. Given the large amount of data stored in database, it is useful to be describe concepts in concise and succinct terms at generalized at multiple levels of abstraction facilities users in examining the general behavior of the data. Given the AB Company database, for example, instead of examining individual customer transactions, sales managers may prefer to view the data generalized to higher levels, such as summarized to higher levels, such as summarized by customer groups according to geographic regions, frequency of purchases per group, and customer income. Such multiple dimensional, multilevel data generalization is similar to multidimensional data analysis in data warehouses.	
	b)	**Attempt any one of the following:**	**(1×6=6)**
	a)	**Describe classes of DSS and categories of the same.**	**6M**
	Ans:	DSS have been classified in different ways as the concept matured with time. As. and when the full potential and possibilities for the field emerged, different classification systems also emerged. Some of the well-known classification models are given below. **According to Donovan and Mad nick (1977) DSS can be classified as:** **1).** Institutional-when the DSS supports ongoing and recurring decisions **2).** Ad hoc-when the DSS supports a one off-kind of decision. Hack thorn and Keen (1981) classified DSS as, **1).** Personal DSS **2).** Group DSS **3).** Organizational DSS Alter (1980) opined that decision support systems could be classified into seven types based on their generic nature of operations. He described the seven types as, **1).** File drawer systems. This type of DSS primarily provides access to data stores/data related items. Examples--ATM Machine, Use the balance to make transfer of funds decisions **2).** Data analysis systems. This type of DSS supports the manipulation of data through the use of specific or generic computerized settings or tools. Examples: Airline Reservation system, use the info to make flight plans **3).** Analysis information systems. This type of DSS provides access to sets of decision oriented databases and simple small models. **4).** Accounting and financial models. This type of DSS can perform 'what if analysis' and calculate the outcomes of different decision paths. Examples: calculate production cost,	***(Classes: 3 marks (any 3), Categories: 3 marks (any 3)***

		make pricing decisions **5).** Representational models. This type of DSS can also perform 'what if analysis' and calculate the outcomes of different decision paths, based on simulated models. **6).** Optimization models. This kind of DSS provides solutions through the use of optimization models which have mathematical solutions. **7).** Suggestion models. This kind of DSS works when the decision to be taken is based on well-structured tasks. Examples: Expert System• Applicant applies for personal loan Modern classification of DSS are, **1. Data driven DSS:** These DSS has file drawer systems, data analysis systems, analysis information systems, data warehousing and emphasizes access to and manipulation of large databases of structured data **2. Model driven:** The underlying model that drives the DSS can come from various disciplines or areas of specialty and might include accounting models, financial models, representation models, optimization models, etc. With model drive DSS emphasize is on access to and manipulation of a model, rather than data, i.e. it uses data and parameters to aid decision makers in analyzing a situation. These systems usually are not data intensive and consequently are not linked to very large databases. **3. Knowledge driven:** These systems provide recommendation and/or suggestion schemes which aid the user in selecting an appropriate alternative to a problem at hand. Knowledge driven DSS are often referred to as management expert systems or intelligent decision support systems. They focus on knowledge and recommends actions to managers based on an analysis of a certain knowledge base. Moreover, it has special problem solving expertise and are closely related to data mining i.e. sifting through large amounts of data to produce contend relationships. **4. Document driven:** These systems help managers retrieve and mange unstructured documents and web pages by integrating a variety of storage and processing technologies to provide complete document retrieval and analysis. It also access documents such as company policies and procedures, product specification, catalogues, corporate historical documents, minutes of meetings, important correspondence, corporate records, etc. and are usually driven by a task-specific search engine. **5. Communication driven:** This breed of DSS is often called group decision support systems (GDSS). They are a special type of hybrid DSS that emphasizes the use of communications and decision models intended to facilitate the solution of problems by decision makers working together as a group. GDSS supports electronic communication, scheduling, document sharing and other group productivity and decision enhancing activities and involves technologies such as two-way interactive video, bulletin boards, e-mail, etc.	
	b)	**Describe need of data warehousing and characteristics of data warehousing.**	**6M**
	Ans:	**Need of Data Warehousing** **1) Advanced query processing**: in most businesses, even the best database systems are bound to either a single server or a handful of servers in a cluster. A data warehouse is	***(Any 3 Need of Data Warehousin***

		a purpose-built hardware solution far more advanced than standard database servers. What this means is a data warehouse will process queries much faster and more effectively, leading to efficiency and increased productivity. **2) Better consistency of data:** developers work with data warehousing systems after data has been received so that all the information contained in the data warehouse is standardized. Only uniform data can be used efficiently for successful comparisons. Other solutions simply cannot match a data warehouse's level of consistency. **3) Improved user access**: a standard database can be read and manipulated by programs like SQL Query Studio or the Oracle client, but there is considerable ramp up time for end users to effectively use these apps to get what they need. Business intelligence and data warehouse end-user access tools are built specifically for the purposes data warehouses are used: analysis, benchmarking, prediction and more. **4) All-in-one**: a data warehouse has the ability to receive data from many different sources, meaning any system in a business can contribute its data. Let's face it: different business segments use different applications. Only a proper data warehouse solution can receive data from all of them and give a business the "big picture" view that is needed to analyze the business, make plans, track competitors and more. **5) Future-proof:** a data warehouse doesn't care where it gets its data from. It can work with any raw information and developers can "massage" any data it may have trouble with. Considering this, you can see that a data warehouse will outlast other changes in the business' technology. For example, a business can overhaul its accounting system, choose a whole new CRM solution or change the applications it uses to gather statistics on the market and it won't matter at all to the data warehouse. Upgrading or overhauling apps anywhere in the enterprise will not require subsequent expenditures to change the data warehouse side. **6) Retention of data history:** end-user applications typically don't have the ability, not to mention the space, to maintain much transaction history and keep track of multiple changes to data. Data warehousing solutions have the ability to track all alterations to data, providing a reliable history of all changes, additions and deletions. With a data warehouse, the integrity of data is ensured. **7) Disaster recovery implications**: a data warehouse system offers a great deal of security when it comes to disaster recovery. Since data from disparate systems is all sent to a data warehouse, that data warehouse essentially acts as another information backup source, Considering the data warehouse will also be backed up, that's now four places where the same information will be stored: the original source, its backup, the data warehouse and its subsequent backup. This is unparalleled information security. **Characteristic of Data Warehousing** **1) Subject Oriented:** Data warehouses are designed to help you analyse data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case makes the data warehouse subject oriented. **2) Integrated:** Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this,	***g: 3 marks, Any 3 Characteristic: 3 marks)***

they are said to be integrated.

3) Non-volatile: Non-volatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyse what has occurred.

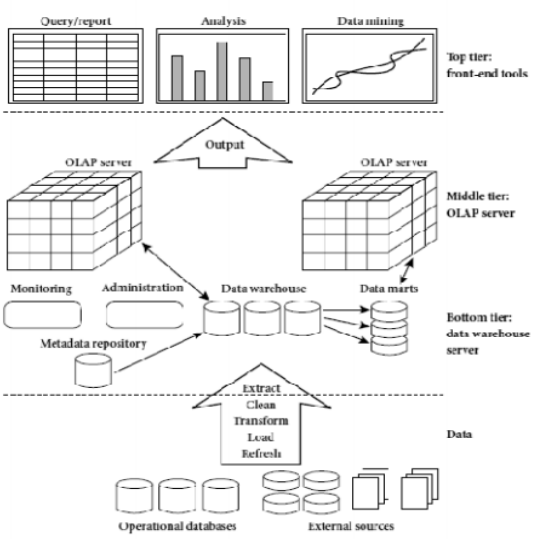
4) Time Variant: In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant. Typically, data flows from one or more online transaction processing (OLTP) databases into a data warehouse on a monthly, weekly, or daily basis. The data is normally processed in a staging file before being added to the data warehouse. Data warehouses commonly range in size from tens of gigabytes to a few terabytes. Usually, the vast majority of the data is stored in a few very large fact tables.

5) Separate: The DW is separate from the operational systems in the company. It gets its data out of these legacy systems.

6) Available: The task of a DW is to make data accessible for the user.

7) Aggregation performance: The data which is requested by the user has to perform well on all scales of aggregation.

8) Consistency: Structural and contents of the data is very important and can only be guarantee by the use of metadata: this is independent from the source and collection date of the data.

2.		**Answer any two of the following:**	**(2×8=16)**
	a)	**Explain with neat block diagram data warehousing and its components functions.**	**8M**
	Ans:		***(Block diagram: 4 marks, Components function: 4 marks)***

1. The bottom tier is a warehouse database server that is almost always a relational database system. Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants). These tools and utilities perform data extraction,

		cleaning, and transformation (e.g., to merge similar data from different Sources into a unified format), as well as load and refresh functions to update the data warehouse. The data are extracted using application program interfaces known as gateways. A gateway is supported by the underlying DBMS and allows client programs to generate SQL code to be executed at a server. Examples of gateways include ODBC (Open Database Connection) and OLEDB (Open Linking and Embedding for Databases) by Microsoft and JDBC (Java Database Connection). **2.** This tier also contains a metadata repository, which stores information about the data warehouse and its contents. The middle tier is an OLAP server that is typically implemented using either (1) a relational OLAP (ROLAP) model, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations; or (2) A multidimensional OLAP (MOLAP) model, that is, a special-purpose server that directly implements multidimensional data and operations. **3.** The top tier is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).	
	b)	**Describe the need of data preprocessing and its techniques. Draw neat block diagrams.**	**8M**
	Ans:	**Need of Data preprocessing:** Data preprocessing converts raw data and signals into data representation suitable for application through a sequence of operations. The objectives of data preprocessing include size reduction of the input space, smoother relationships, data normalization, noise reduction, and feature extraction. Several data preprocessing algorithms, such as data values averaging, input space reduction, and data normalization, will be briefly discussed in this chapter. Computer programs for data preprocessing are also provided. **Techniques in Data preprocessing:** **1.** Data Integration: Data Integration is a data preprocessing technique that merges the data from multiple heterogeneous data sources into a coherent data store. Data integration may involve inconsistent data and therefore needs data cleaning. **2.** Data Cleaning: Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse. **3.** Data Selection: Data Selection is the process where data relevant to the analysis task are retrieved from the database. Sometimes data transformation and consolidation are performed before the data selection process. **4.** Clusters: Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters. **5.** Data Transformation: Data is transformed or consolidated into forms appropriate for mining, by performing summary or aggregation operations.	***(Need: 2 marks, Techniques: 4 marks, Block diagram: 2 marks)***

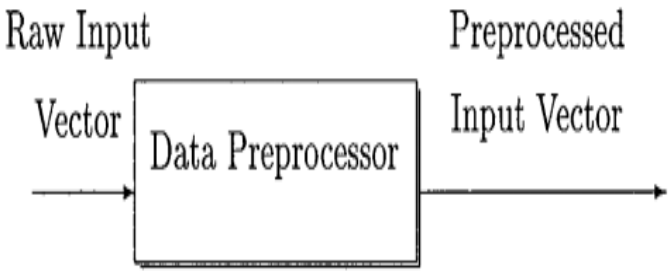


Fig: Block diagram of Data Preprocessing

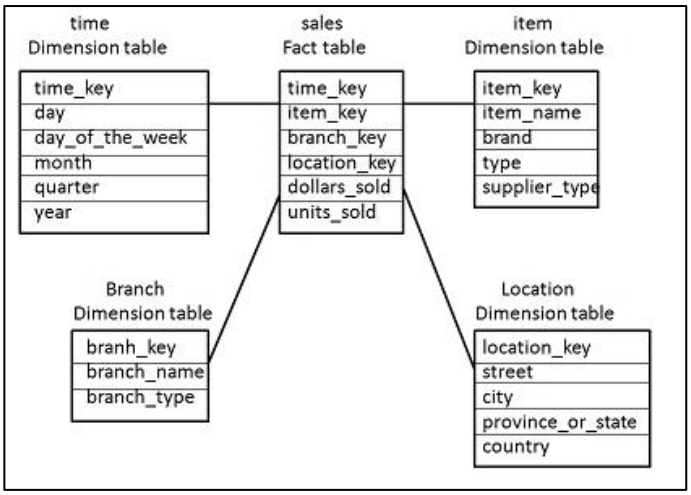
c) Describe following schema's for multidimensional data base 1) Star 2) Snow flakes. 8M

Ans:

Star Schema:

- In star schema each dimension is represented with only one dimension table.
- This dimension table contains the set of attributes.
- In the following diagram we have shown the sales data of a company with respect to the four dimensions namely, time, item, branch and location.

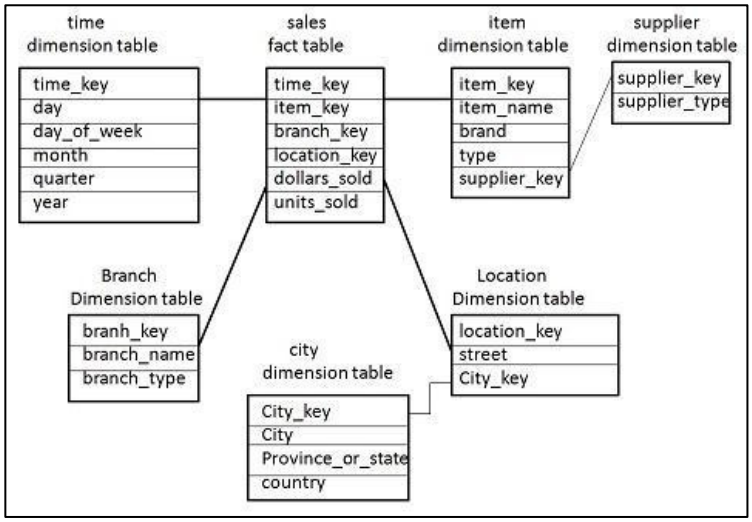
(Star: 4 marks, Snowflakes: 4 marks)



- There is a fact table at the center. This fact table contains the keys to each of four dimensions.
- The fact table also contain the attributes namely, dollars sold and units sold.

Snowflake Schema:

- In Snowflake schema some dimension tables are normalized.
- Dimensions with hierarchies can be decomposed into a snowflake structure when you want to avoid joins to big dimension tables when you are using an aggregate of the fact table.
- The normalization split up the data into additional tables.
- Unlike Star schema the dimensions table in snowflake schema is normalized for example the item dimension table in star schema is normalized and split into two dimension tables namely, item and supplier table.



- Therefore now the item dimension table contains the attributes item key, item name, type, brand, and supplier-key.
- The supplier key is linked to supplier dimension table. The supplier dimension table contains the attributes supplier key, and supplier type.

3.		**Answer any four of the following:**	**(4×4=16)**
	a)	**Describe model management for DSS.**	**4M**
	Ans:	DSSs include knowledge-based systems. A properly designed DSS is an interactive software based system intended to help decision makers compile useful information from a combination of raw data, documents, and personal knowledge, or business models to identify and solve problems and make decisions. **Components:** Design of a drought mitigation decision support system three fundamental components of a DSS architecture are: **1.** The database (or knowledge base), **2.** The model (i.e., the decision context and user criteria), and **3**. The user interface (allows you to communicate with the DSS). **Model Management Component:** The model management component consists of both the Decision Support System models and the Decision Support System model management system. A model is a representation of some event, fact, or situation. As it is not always practical, or wise, to experiment with reality, people build models and use them for experimentation. Models can take various forms. Businesses use models to represent variables and their relationships. For example, you would use a statistical model called analysis of variance to determine whether newspaper, TV, and billboard advertising are equally effective in increasing sales. Decision Support Systems help in various decision-making situations by utilizing models that allow you to analyze information in many different ways. The models you use in a Decision Support System depend on the decision you are making and, consequently, the kind of analysis you require. For example, you would use what-if analysis to see what effect the change of one or more variables will have on other variables, or optimization to find the most profitable solution given operating restrictions and limited resources. Spreadsheet software such as excel can be used as a	***(Explanation : 4 marks)***

		Decision Support System for what-if analysis. The model management system stores and maintains the Decision Support System's models. Its function of managing models is similar to that of a database management system. The model management component cannot select the best model for you to use for a particular problem that requires your expertise but it can help you create and manipulate models quickly and easily.	
	b)	**Explain DSS and its implementation in business organization.**	**4M**
	Ans:	Decision support systems are interactive software-based systems intended to help managers in decision making by accessing large volume of information generated from various related information systems involved in organizational business processes, like, office automation system, transaction processing system etc. DSS uses the summary information, exceptions, patterns and trends using the analytical models. Decision Support System helps in decision making but does not always give a decision itself. The decision makers compile useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions. **Programmed and Non-programmed Decisions:** There are two types of decisions - programmed and non-programmed decisions. Programmed decisions are basically automated processes, general routine work, where: • These decisions have been taken several times • These decisions follow some guidelines or rules Non-programmed decisions occur in unusual and non-addressed situations, so: • It would be a new decision • There will not be any rules to follow • These decisions are made based on available information • These decisions are based on the manger's discretion, instinct, perception and judgment. Decision support systems generally involve non-programmed decisions. Therefore, there will be no exact report, content or format for these systems.	***(Explanation : 4 marks)***
	c)	**Describe data reduction techniques.**	**4M**
	Ans:	Imagine that you have selected data from the *All Electronics* data warehouse for analysis. The data set will likely be huge! Complex data analysis and mining on huge amounts of data can take a long time, making such analysis impractical or infeasible. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following: **1.** Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube. **2.** Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed. **3.** Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. **4.** Numerosity reduction ,where the data are replaced or estimated by alternative, smaller	***(Explanation : 4 marks)***

		data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms. **5.** Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and concept hierarchy generation are powerful tools for data mining, in that they allow the mining of data at multiple levels of abstraction.	
	d)	**Define OLAP and why it is required for data warehousing.**	**4M**
	Ans:	OLAP (online analytical processing) is a function of business intelligence software that enables a user to easily and selectively extract and view data from different points of view. OLAP technology is a vast improvement over traditional relational database management systems (RDBMS). Relational databases, which have a two-dimensional structure, do not allow the multidimensional data views that OLAP provides. Traditionally used as an analytical tool for marketing and financial reporting, OLAP is now viewed as a valuable tool for any management system that needs to create a flexible decision support system. Today's work environment is characterized by flatter organizations that need to be able to adapt quickly to changing conditions. Managers need the tools that will allow them to make quick, intelligent decisions on the fly. Making the wrong decision or taking too long to make it can affect the competitive position of an organization. OLAP provides the multidimensional capabilities that most organizations need today. By using a multidimensional data store, also known in the industry as a hypercube, OLAP allows the end user to analyze data along the axes of their business. The two most common forms of analysis that most businesses use are called "slice and dice" and "drill down".	***(Define: 1 mark and requirement: 3 marks)***
	e)	**Give benefits of data warehousing.**	**4M**
	Ans:	Benefits from a successful implementation of a data warehouse include: • Enhanced Business Intelligence Insights will be gained through improved information access. Managers and executives will be freed from making their decisions based on limited data. Decisions that affect the strategy and operations of organizations will be based upon credible facts and will be backed up with evidence and actual organizational data. • Increased Query and System Performance The data warehouse is built for analysis and retrieval of data rather than efficient upkeep of individual records (i.e. transactions). Further, the data warehouse allows for a large system burden to be taken off the operational environment and effectively distributes system load across an entire organization's technology infrastructure. • Business Intelligence from Multiple Sources For many organizations, enterprise information systems are comprised of multiple subsystems, physically separated and built on different platforms. Moreover, merging data from multiple disparate data sources is a common need when conducting business intelligence. To solve this problem, the data warehouse performs integration of existing disparate data sources and makes them accessible in one place. • Timely Access to Data The data warehouse enables business users and decision makers to have access to data from many different sources as they need to have access to the data. Additionally, business users will spend little time in the data retrieval process. Scheduled	***(List of benefits: 1 mark, Explanation: 3 marks)***

		data integration routines, known as ETL, are leveraged within a data warehouse environment. • Enhanced Data Quality and Consistency A data warehouse implementation typically includes the conversion of data from numerous source systems and data files and transformation of the disparate data into a common format. Data from the various business units and departments is standardized and the inconsistent nature of data from the unique source systems is removed. Moreover, individual business units and departments including sales, marketing, finance, and operations, will start to utilize the same data repository as the source system for their individual queries and reports. Thus each of these individual business units and departments will produce results that are consistent with the other business units within the organization. • Historical Intelligence Data warehouses generally contain many years' worth of data that can neither be stored within nor reported from a transactional system. Typically transactional systems satisfy most operating reporting requirements for a given time-period but without the inclusion of historical data. In contrast, the data warehouse stores large amounts of historical data and can enable advanced business intelligence including time-period analysis, trend analysis, and trend prediction. The advantage of the data warehouse is that it allows for advanced reporting and analysis of multiple time-periods. • High Return on Investment Return on investment (ROI) refers to the amount of increased revenue or decreased expenses a business will be able to realize from any project or investment of capital. Subsequently, implementations of data warehouses and complementary business intelligence systems have enabled business to generate higher amounts of revenue and provide substantial cost savings.	
4.	**a)**	**Answer any three of the following:**	**(3×4=12)**
	a)	**State and explain mining to world wide web.**	**4M**
	Ans:	Data mining refers to extracting or "mining" knowledge from large amounts of data. Mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. The World Wide Web contains the huge information such as hyperlink information, web page access info, education etc that provide rich source for data mining. The basic structure of the web page is based on Document Object Model (DOM). The DOM structure refers to a tree like structure. In this structure the HTML tag in the page corresponds to a node in the DOM tree. We can segment the web page by using predefined tags in HTML. The HTML syntax is flexible therefore; the web pages do not follow the W3C specifications. Not following the specifications of W3C may cause error in DOM tree structure. The DOM structure was initially introduced for presentation in the browser not for description of semantic structure of the web page. The DOM structure cannot correctly identify the semantic relationship between different parts of a web page.	***(Term: 1 mark, Explanation: 3 marks)***
	b)	**State and explain issues regarding classification and predictions.**	**4M**
	Ans:	**Preparing the Data for Classification and Prediction:** The following preprocessing steps may be applied to the data in order to help improve the accuracy, efficiency, and scalability of the classification or prediction process. **Data Cleaning:** This refers to the preprocessing of data in order to remove or reduce noise	***(State: 1 mark and any two issues: 1 ½***

		(by applying smoothing techniques) and the treatment of missing values (e.g., by replacing a missing value with the most commonly occurring value for that attribute, or with the most probable value based on statistics.) Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help reduce confusion during learning. **Relevance Analysis:** Many of the attributes in the data may be irrelevant to the classification or prediction task. For example, data recording the day of the week on which a bank loan application was filed is unlikely to be relevant to the success of the application. Furthermore, other attributes may be redundant. Hence, relevance analysis may be performed on the data with the aim of removing any irrelevant or redundant attributes from the learning process. In machine learning, this step is known as feature selection. Including such attributes may otherwise slow down, and possibly mislead, the learning step. Ideally, the time spent on relevance analysis, when added to the time spent on learning from the resulting "reduced" feature subset should be less than the time that would have been spent on learning from the original set of features. Hence, such analysis can help improve classification efficiency and scalability. **Data Transformation:** The data can be generalized to higher – level concepts. Concept hierarchies may be used for this purpose. This is particularly useful for continuous – valued attributes. For example, numeric values for the attribute income may be generalized to discrete ranges such as low, medium, and high. Similarly, nominal – valued attributes like street, can be generalized to higher – level concepts, like city. Since generalization compresses the original training data, fewer input / output operations may be involved during learning. The data may also be normalized, particularly when neural networks or methods involving distance measurements are used in the learning step. **Normalization** involves scaling all values for a given attribute so that they fall within a small specified range, such as – 1.0 to 1.0, or 0.0 to 1.0. In methods that use distance measurements, for example, this would prevent attributes with initially large ranges (like, say, income) from outweighing attributes with initially smaller ranges (such as binary attributes). **Comparing Classification Methods:** Classification and prediction methods can be compared and evaluated according to the following criteria: **Predictive Accuracy:** This refers to the ability of the model to correctly predict the class label of new or previously unseen data. **Speed:** This refers to the computation costs involved in generating and using the model. **Robustness:** This is the ability of the model to make correct predictions given noisy data or data with missing values. **Scalability:** This refers to the ability to construct the model efficiently given large amount of data. **Interpretability:** This refers to the level of understanding and insight that is provided by the model.	***marks)***
	c)	**Explain mining text database.**	**4M**
	Ans:	A substantial portion of the available information is stored in text databases (or document databases), which consist of large collections of documents from various sources, such as news articles, research papers, books, digital libraries, e-mail messages, and Web pages.	***(Explanation : 4 marks)***

		Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web (which can also be viewed as a huge, interconnected, dynamic text database). Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases. Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain structured fields, such as title, authors, publication date, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modelling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining. Text Mining Approaches: There are many approaches to text mining, which can be classified from different perspectives, based on the inputs taken in the text mining system and the data mining tasks to be performed. In general, the major approaches, based on the kinds of data they take as input, are: **1.** The keyword-based approach, where the input is a set of keywords or terms in the documents, **2.** The tagging approach, where the input is a set of tags, and **3.** The information-extraction approach, which inputs semantic information, such as events, facts, or entities uncovered by information extraction.	
	d)	**List four major applications of data mining in business.**	**4M**
	Ans:	Here is the list of 14 other important areas where data mining is widely used: **Future Healthcare:** Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse. **Market Basket Analysis:** Market basket analysis is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done. Education There is	***(Any four application: 1 mark Each)***

a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

Manufacturing Engineering: Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

CRM: Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyze the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

Fraud Detection: Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

Intrusion Detection: Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

Lie Detection: Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data samples collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

Customer Segmentation: Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the

		customers. Market is always about retaining the customers. Data mining allows to find a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction. **Financial Banking:** With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a profitable customer. **Corporate Surveillance:** Corporate surveillance is the monitoring of a person or group's behavior by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails. **Research Analysis:** History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of the data. **Criminal Investigation:** Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. This information can be used to perform crime matching process. **Bio Informatics:** Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.	
	b)	**Answer any one of the following:**	**(1×6=6)**
	a)	**Describe the method of summarization based on characterization.**	**6M**
	Ans:	For data preprocessing to be successful, it is essential to have an overall picture of your data. Descriptive data summarization techniques can be used to identify the typical	***(Explanation***

		properties of your data and highlight which data values should be treated as noise or outliers. For many data preprocessing tasks, users would like to learn about data characteristics regarding both central tendency and dispersion of the data. Measures of central tendency include mean, median, mode, and midrange, while measures of data dispersion include quartiles, interquartile range (IQR), and variance. These descriptive statistics are of great help in understanding the distribution of the data. Such measures have been studied extensively in the statistical literature. From the data mining point of view, we need to examine how they can be computed efficiently in large databases. In particular, it is necessary to introduce the notions of distributive measure, algebraic measure, and holistic measure. Knowing what kind of measure we are dealing with can help us choose an efficient implementation for it. **Measuring the Central Tendency:** The most common and most effective numerical measure of the "center" of a set of data is the (arithmetic) mean. Let x1;x2; : : : ;xN be a set of N values or observations, such as for some attribute, like salary. The mean of this set of values is $$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$ This corresponds to the built-in aggregate function, average (avg() in SQL), provided in relational database systems. A distributive measure is a measure (i.e., function) that can be computed for a given data set by partitioning the data into smaller subsets, computing the measure for each subset, and then merging the results in order to arrive at the measure's value for the original (entire) data set. Both sum() and count() are distributive measures because they can be computed in this manner. Other examples include max () and min (). An algebraic measure is a measure that can be computed by applying an algebraic function to one or more distributive measures. Hence, average (or mean ()) is an algebraic measure because it can be computed by sum()/count(). A holistic measure is a measure that must be computed on the entire data set as a whole. It cannot be computed by partitioning the given data into subsets and merging the values obtained for the measure in each subset. The median is an example of a holistic measure.	***: 6 marks)***
	b)	**List and explain application of knowledge discovery techniques.**	**6M**
	Ans:	**List of Knowledge discovery:** Associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. **1. Association analysis:** Suppose, as a marketing manager of All Electronics, you would like to determine which items are frequently purchased together within the same transactions. An example of such a rule, mined from the All Electronics transactional database, is buys(X ,—computer‖) ⇒ buys(X , —software‖) [support = 1%, confidence = 50%] where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together. This association rule involves a single	***(List:1 mark , Explanation: 5 marks)***

attribute or predicate (i.e., buys) that repeats. Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as ―computer⇒ software [1%, 50%]‖.Suppose, instead, that we are given the All Electronics relational database relating to purchases. A data mining system may find association rules like age(X , ―20...29‖) ∧ income(X , ―20K...29K‖) ⇒ buys(X , ― CD player‖)[support = 2%, confidence = 60%] The rule indicates that of the All Electronics customers under study, 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player at All Electronics. There is a 60% probability that a customer in this age and income group will purchase a CD player.

2. Classification: Is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Example: Suppose, as sales manager of All Electronics, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: good response, mild response, and no response. You would like to derive a model for each of these three classes based on the descriptive features of the items, such as price, brand, place made, type, and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

3. Clustering analyzes: data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the infraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Example: Cluster analysis can be performed on All Electronics customer data in order to identify Homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.

4. Outlier Analysis: A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

Example: Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account. Outlier values may also be detected with respect to the location and type of purchase, or the purchase frequency.

5. Evolution Analysis: Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Example: Suppose that you have the major stock market (time-series) data of the last several years available from the New York Stock Exchange and you would like to invest in shares of high-tech industrial companies. A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies. Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

5.		**Answer any two of the following:**	**(2×8=16)**
	a)	**Describe with example the apropri alogorithm.**	**8M**
	Ans:	Apriori is a seminal algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties Apriori employs an iterative approach known as a level-wise search, where k item sets are used to explore (k + 1)-item sets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2 the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-item sets can be found. The finding of each Lk requires one full scan of the database. Once the frequent item sets from transactions in a database D have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence). This can be done using Equation for confidence, which we show again here for completeness: $$confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}.$$ **Input:**D, a database of transactions; Min_sup, the minimum support count threshold **Output:** L, frequent itemsets in D	***(Explanation : 4 marks, Example: 4 marks)***

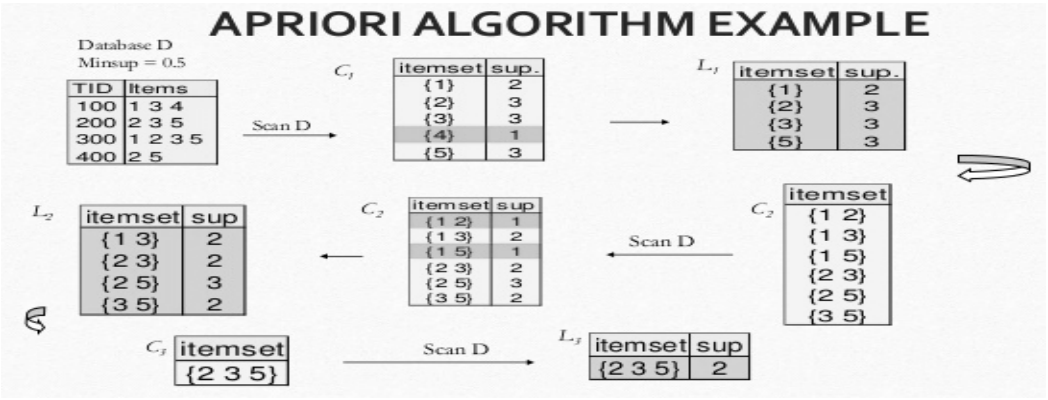
```
Method:
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)      Ck = apriori_gen(Lk−1);
(4)      for each transaction t ∈ D { // scan D for counts
(5)         Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)         for each candidate c ∈ Ct
(7)            c.count++;
(8)      }
(9)      Lk = {c ∈ Ck | c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;

procedure apriori_gen(Lk−1:frequent (k − 1)-itemsets)
(1)   for each itemset l1 ∈ Lk−1
(2)      for each itemset l2 ∈ Lk−1
(3)         if (l1[1] = l2[1]) ∧ (l1[2] = l2[2]) ∧ ... ∧ (l1[k − 2] = l2[k − 2]) ∧ (l1[k − 1] < l2[k − 1]) then {
(4)            c = l1 ⋈ l2; // join step: generate candidates
(5)            if has_infrequent_subset(c, Lk−1) then
(6)               delete c; // prune step: remove unfruitful candidate
(7)            else add c to Ck;
(8)         }
(9)      return Ck;

procedure has_infrequent_subset(c: candidate k-itemset;
         Lk−1: frequent (k − 1)-itemsets); // use prior knowledge
(1)   for each (k − 1)-subset s of c
(2)      if s ∉ Lk−1 then
(3)         return TRUE;
(4)      return FALSE;
```

The Apriori property is used in the algorithm by a two-step process consisting of **join** and **prune** actions.

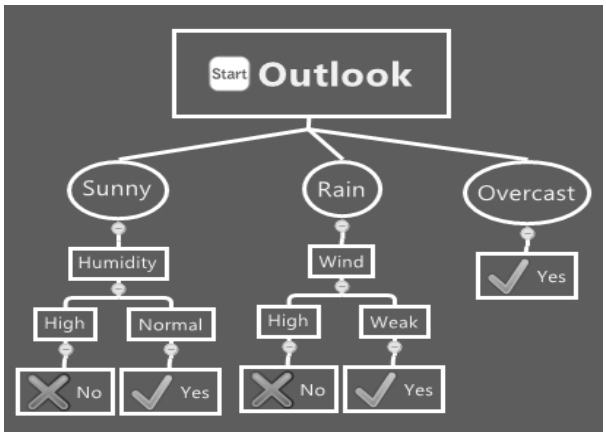
Apriori Example:



	b)	**List all mining techniques and explain any one.**	**8M**
	Ans:	There are several major data mining techniques have been developing and using in data	***(List: 2***

		mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections. **Association:** Association is one of the best-known data mining techniques. In association, a pattern is discovered based on a relationship between items in the same transaction. That's the reason why association technique is also known as *relation technique*. The association technique is used in *market basket analysis* to identify a set of products that customers frequently purchase together. Retailers are using association technique to research customer's buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for customer and increase sales. **Classification:** Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups. **Clustering:** Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library. **Prediction:** The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction. **Sequential Patterns:** Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.	***marks, Explanation of any one: 6 marks)***

Decision trees: The A decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it. For example, We use the following decision tree to determine whether or not to play tennis:



Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal. We often combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

	c)	**List all mining associations rule and explain any one of it.**	**8M**
	Ans:	Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a customer buys a dozen eggs, he is 80% likely to also purchase milk. “It is often unrealistic and inefficient for data mining systems to generate all possible patterns. Instead, user provided constraints and interestingness measures should be used to focus the search. For some mining tasks, such as association, this is often sufficient to ensure the completeness of the algorithm. Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining. In general, association rule mining can be viewed as a two-step process: **1. Find all frequent itemsets:** By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min_sup. **2. Generate strong association rules from the frequent itemsets:** By definition, these rules must satisfy minimum support and minimum confidence. A major challenge in mining	***(List: 2 marks, Explanation: 6 marks)***

frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support. (min_sup) threshold, especially when min sup is set low. To overcome this difficulty, we introduce the concepts of closed frequent itemset and maximal frequent itemset. An itemset X is **closed** in a data set D if there exists no proper super-itemset Y^5 such that Y has the same support count as X inD. An itemset X is a **closed frequent itemset** in set D if X is both closed and frequent in D. An itemset X is a **maximal frequent itemset** (or **max-itemset**) in a data set D if X is frequent, and there exists no super-itemset Y such that X C Y and Y is frequent in D.

Association rule mining: consists of first finding **frequent itemsets** (sets of items, such as A and B, satisfying a minimum support threshold, or percentage of the task relevant tuples), from which **strong** association rules are generated. These rules also satisfy a minimum confidence threshold (a prespecified probability of satisfying B under the condition that A is satisfied). Associations can be further analyzed to uncover **correlation rules**, which convey statistical correlations between itemsets A and B. Many efficient and scalable algorithms have been developed for **frequent itemset mining**, from which association and correlation rules can be derived. These algorithms can be classified into three categories:

(1) Apriori-like algorithms,
(2) frequent pattern growth–based algorithms such as FP-growth, and
(3) Algorithms that use the vertical data format.
Apriori is a seminal algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties Apriori employs an iterative approach known as a level-wise search, where k itemsets are used to explore (k + 1)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted L1. Next, L1 is used to find L2 the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent k-itemsets can be found. The finding of each Lk requires one full scan of the database. Once the frequent itemsets from transactions in a database D have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence).
Frequent pattern growth: is a method of mining frequent itemsets without candidate generation. It constructs a highly compact data structure (an FP-tree) to compress the original transaction database. Rather than employing the generate-and-test strategy of Apriori-like methods, it focuses on frequent pattern (fragment) growth, which avoids costly candidate generation, resulting in greater efficiency.
Mining frequent itemsets using the vertical data format (Eclat): is a method that transforms a given data set of transactions in the horizontal data format of TIDitemset into the vertical data format of item-TID set. It mines the transformed data set by TID set intersections based on the Apriori property and additional optimization techniques such as diffset.

6.		**Answer any four of the following:**	**(4×4=16)**
	a)	**State any six needs of data mining.**	**4M**
	Ans:	**a) Customer Profiling**: Data mining helps determine what kind of people buy what kind of products. **b)** Identifying Customer Requirements: Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers. **c) Cross Market Analysis**: Data mining performs Association/correlations between product sales. **d) Target Marketing**: Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc. **e) Determining Customer purchasing pattern**: Data mining helps in determining customer purchasing pattern. **f) Providing Summary Information**: Data mining provides us various multidimensional summary reports. **g) Finance Planning and Asset Evaluation**: It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets. **h) Resource Planning**: It involves summarizing and comparing the resources and spending. **Competition –** It involves monitoring competitors and market directions. Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms	***(Any six Needs : 4 marks)***
	b)	**Define metadata. How it will be classified according to need of organization?**	**4M**
	Ans:	**Metadata** are data about data. When used in a data warehouse, metadata are the data that define warehouse objects. Metadata are created for the data names and definitions of the given warehouse. Additional metadata are created and captured for time stamping any extracted data, the source of the extracted data, and missing fields that have been added by data cleaning or integration processes. A metadata repository should contain the following: **i)** A description of the data warehouse structure, which includes the warehouse schema View, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents. **ii)** Operational metadata, which include data lineage (history of migrated data and the Sequence of transformations applied to it), currency of data (active, archived, or	***(Definition: 1 mark, Classification: 3 marks)***

		purged), and monitoring information (warehouse usage statistics, error reports, and audit trails). **iii)** The algorithms used for summarization, which include measure and dimension Definition algorithms, data on granularity, partitions, subject areas, aggregation, summarization, and predefined queries and reports. **iv)** Mapping from the operational environment to the data warehouse, which includes Source databases and their contents, gateway descriptions, data partitions, data extraction, cleaning, transformation rules and defaults, data refresh and purging rules, and security (user authorization and access control). **v)** Data related to system performance, which include indices and profiles that improve data access and retrieval performance, in addition to rules for the timing and scheduling of refresh, update, and replication cycles. **vi)** Business metadata, which include business terms and definitions, data ownership Information and charging policies.	
	c)	**Describe mining descriptive statistical measures in large database.**	**4M**
	Ans:	Class description can be explained with respect to the terms of popular Measures, such as count, sum, and average. Relational database systems provide five Built-in aggregate functions: count ( ), sum ( ), max ( ), and min ( ). These Functions can also be computed efficiently (in incremental and distributed manners) in data cubes. Thus, there is no problem in including these aggregate functions as basic measures in the descriptive mining of multidimensional data. For many data mining tasks, however, users would like to learn more data characteristics regarding both central tendency and data dispersion. Measures of central tendency include mean, median, mode, and midrange, while measures of data dispersion include quartiles, outliers, and variance. These descriptive statistics are of great help in Understanding the distribution of the data. Such measures have been studied extensively In the statistical literature. From the data mining point of view, we need to examine how They can be computed efficiently in large multidimensional databases.	***(Description : 4 marks)***
	d)	**How data mining algorithms can be implemented in various applications of data mining? Justify your answer.**	**4M**
	Ans:	An algorithm in data mining (or machine learning) is a set of heuristics and calculations that creates a model from data. To create a model, the algorithm first analyzes the data you provide, looking for specific types of patterns or trends. The algorithm uses the results of this analysis over much iteration to find the optimal parameters for creating the mining model. These parameters are then applied across the entire data set to extract actionable patterns and detailed statistics.The mining model that an algorithm creates from your data can take various forms, including: • A set of clusters that describe how the cases in a dataset are related. • A decision tree that predicts an outcome, and describes how different criteria affect that outcome. • A mathematical model that forecasts sales. • A set of rules that describe how products are grouped together in a transaction, and the probabilities that products are purchased together. Choosing the best algorithm to use for a specific analytical task can be a challenge. While you can use different algorithms to perform the same business task, each algorithm produces a different result, and some algorithms can produce more than one type of result. **Choosing an Algorithm by Type**	***(Explanation: 2 marks, Justification : 2 marks)***

SQL Server Data Mining includes the following algorithm types:

- **Classification algorithms** predict one or more discrete variables, based on the other attributes in the dataset.
- **Regression algorithms** predict one or more continuous numeric variables, such as profit or loss, based on other attributes in the dataset.
- **Segmentation algorithms** divide data into groups, or clusters, of items that have similar properties.
- **Association algorithms** find correlations between different attributes in a dataset. The most common application of this kind of algorithm is for creating association rules, which can be used in a market basket analysis.
- **Sequence analysis algorithms** summarize frequent sequences or episodes in data, such as a series of clicks in a web site, or a series of log events preceding machine maintenance.

Choosing an Algorithm by Task

Some examples are:

Example of Task	Algorithm
Categorize patient outcomes and explore related factors.	Neural Network Algorithm
Forecast next year's sales.	Time Series Algorithm
Perform clickstream analysis of a company's Web site.	Sequence Clustering Algorithm