

Acuity Educare

BUSINESS INTELLIGENCE SEM : VI

SEM VI : UNIT 1 to 5



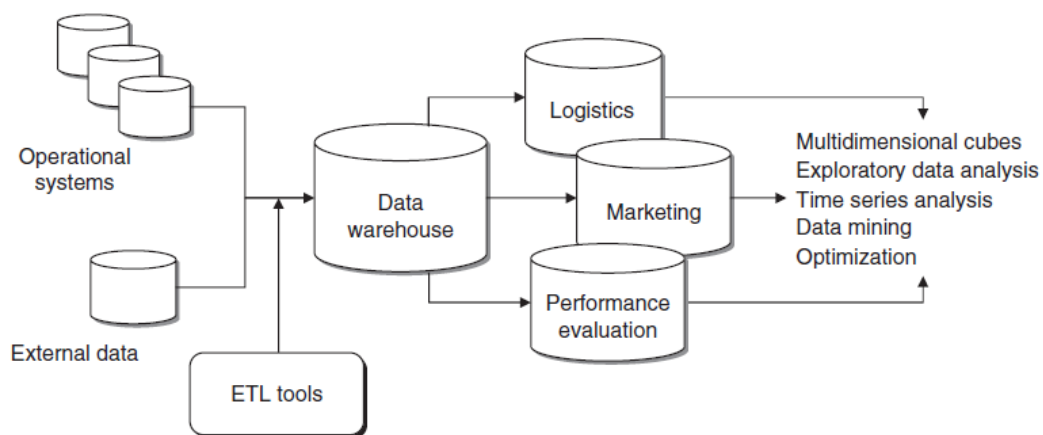
607A, 6th floor, Ecstasy business park, city of joy, JSD
road, mulund (W) | 8591065589/022-25600622



UNIT 1

1. What is business intelligence? Write its advantages.
2. What is business intelligence? Explain architecture of the business intelligence.
Or
Describe the architecture of business intelligence.
3. Explain Data, Information and Knowledge. ? Why effective and timely decision is important?
4. Explain different phases in development business intelligence system.
5. What are the ethics of business Intelligence?
6. Describe the role of mathematical model.
7. Define system. Explain how system can be characterized.
8. What is decision support system (DSS)? What are the factors that affect the degree of success of the DSS?
9. Explain classification of decisions according to their nature and scope.
10. Describe different phases in the development of a decision support systems (DSS).
11. Explain the phases of decision-making process system? Explain types of decision
12. Explain the major potential advantage derived from adoption of a DSS.

1.	What is business intelligence? Write its advantages.
	<ul style="list-style-type: none"> ➤ The term Business Intelligence (BI) refers to technologies, applications and practices for the collection, integration, analysis, and presentation of business information. The main reason behind Business Intelligence is to provide better business decision making. ➤ These systems are data-driven Decision Support Systems (DSS). Business Intelligence is sometimes used interchangeably with briefing books, report and query tools and executive information systems. It is also called as a set of mathematical model and analysis methodology which is very useful for decision making process which are complex. ➤ Large amount of data can be easily accessed by individuals and organizations because of numerous internet connections and low data storage technologies. ➤ Transactions are commercial, financial and administrative, making the data heterogeneous in origin, content and representation. Emails, texts and hypertexts, and the results of clinical tests, are a few examples. ➤ Their accessibility opens various scenarios and opportunities, and raises a rather important question: is it possible to convert such data into information and knowledge that can then be used by decision makers to assist and improve the operation of enterprises and of public administration?
2.	What is business intelligence? Explain architecture of the business intelligence. Or Describe the architecture of business intelligence.
	<p>Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make effective and timely decisions.</p> <p>The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way.</p> <p>If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyse a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions.</p> <p>We may therefore conclude that the major advantage deriving from the adoption of a business intelligence system is found in the increased effectiveness of the decision-making process</p> <p>Business intelligence architectures</p> <p>The architecture of a business intelligence system, depicted in below given figure which includes three major components.</p>



1.Data sources. In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers. Generally speaking, a major effort is required to unify and integrate the different data sources,

2.Data warehouses and data marts. Using extraction and transformation tools known as extract, transform, load (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as data warehouses and data marts

Business intelligence methodologies. Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented. This data is extracted to provide input to mathematical model and support decision makers.

1. Time series analysis;
2. Inductive learning models for data mining;
3. Optimization models.

The description of the upper tiers:

3. Data exploration:

This is the third level called as Data exploration. Data exploration is an informative search which is used by data consumers to form real and true analysis from the information collected. Data Exploration is about describing the data by means of statistical and visualization techniques.

We explore data in order to bring important aspects of that data into focus for further analysis. Often, data is gathered in a non-rigid or controlled manner in large bulks. For true analysis, this unorganized bulk of data needs to be narrowed down. This is where data exploration is used to analyze the data and information from the data to form further analysis.

Data often converges in a central warehouse called a data warehouse. This data can come from various sources using various formats.

Relevant data is needed for tasks such as statistical reporting, trend spotting and pattern spotting. Data exploration is the process of gathering such relevant data.

4. Data mining:

The fourth level is data mining. Data mining technique has to be chosen based on the type of business and the type of problem your business faces.

A generalized approach has to be used to improve the accuracy and cost effectiveness of using data mining techniques

5. Optimization:

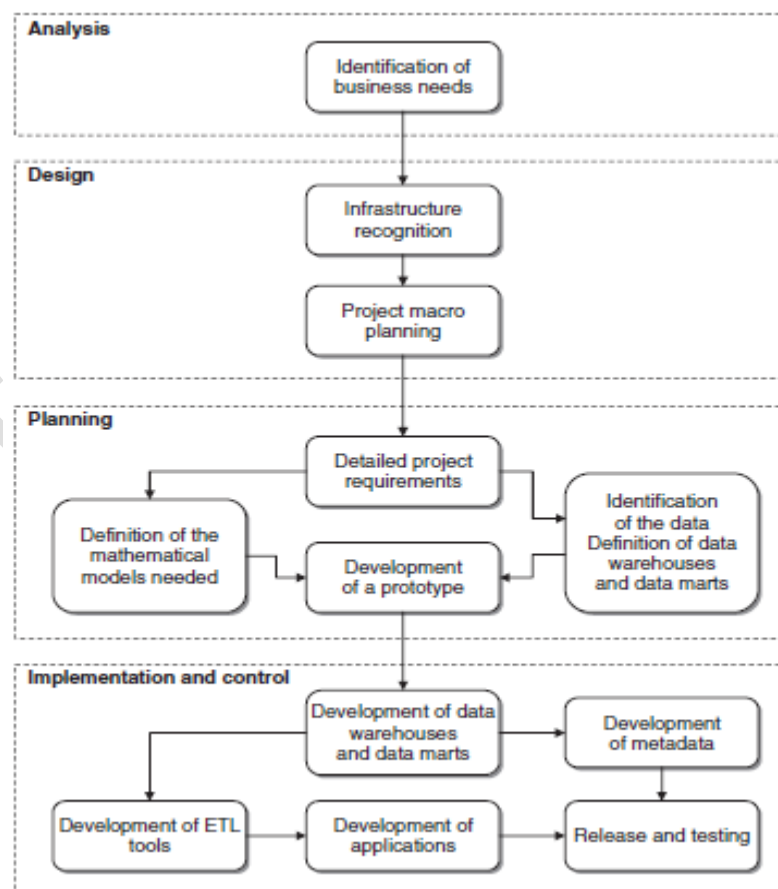
	<p>If we go one level on top we get optimization models which allow us to select best solutions among all other alternative.</p> <p>6. Decisions: The top most level of the pyramid is the decision where we need to select best alternative for decision making process. When business intelligence methodology is successfully adopted it helps to make decision.</p>
3.	Explain Data, Information and Knowledge. ? Why effective and timely decision is important?
	<p>Data: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. For example, for a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions.</p> <p>Information: Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain. For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over ₹100 per week, or the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data.</p> <p>Knowledge: Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems. For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business. The knowledge extracted in this way will eventually lead to actions aimed at solving the problem detected, for example by introducing a new free home delivery service for the customers residing in that specific area. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as knowledge management.</p> <p>Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long- or short-term effects and involve people and roles at various hierarchical levels. The ability of these knowledge workers to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization.</p> <p>The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make effective and timely decisions.</p> <p>Effective decisions: The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way. Indeed, turning to formal analytical methods forces decision makers to explicitly describe both the criteria for evaluating alternative choices and the mechanisms regulating the problem under investigation. Furthermore, the ensuing in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.</p>

Timely decisions: Enterprises operate in economic environments characterized by growing levels of competition and high dynamism. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyse a larger number of alternative actions, achieve more accurate conclusions and reach effective and timely decisions. We may therefore conclude that the major advantage deriving from the adoption of a business intelligence system is found in the increased effectiveness of the decision-making process.

4. Explain different phases in development business intelligence system.

Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. The development of a business intelligence system can be assimilated to a project, with a specific final objective, expected development times and costs, and the usage and coordination of the resources needed to perform planned activities. Below given figure shows the typical development cycle of a business intelligence architecture.



Analysis. During the first phase, the needs of the organization relative to the development of a business intelligence system should be carefully identified. This preliminary phase is generally conducted through a series of interviews of knowledge workers performing different roles and activities within the organization. It is necessary to clearly describe the general objectives and priorities of the project, as well as to set out the costs and benefits deriving from the development of the business intelligence system.

Design. The second phase includes two sub-phases and is aimed at deriving a provisional plan of the overall architecture. First, it is necessary to make an assessment of the existing information infrastructures. Moreover, the main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

Planning. The planning stage includes a sub-phase where the functions of the business intelligence system are defined and described in greater detail. Subsequently, existing data as well as other data that might be retrieved externally are assessed. This allows the information structures of the business intelligence architecture, which consist of a central data warehouse and possibly some satellite data marts, to be designed. Simultaneously with the recognition of the available data, the mathematical models to be adopted should be defined, ensuring the availability of the data required to feed each model and verifying that the efficiency of the algorithms to be utilized will be adequate for the magnitude of the resulting problems. Finally, it is appropriate to create a system prototype, at low cost and with limited capabilities, in order to uncover beforehand any discrepancy between actual needs and project specifications.

Implementation and control. The last phase consists of five main sub-phases. First, the data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system. In order to explain the meaning of the data contained in the data warehouse and the transformations applied in advance to the primary data, a metadata archive should be created. Moreover, ETL procedures are set out to extract and transform the data existing in the primary sources, loading them into the data warehouse and the data marts.

5. What are the ethics of business Intelligence?

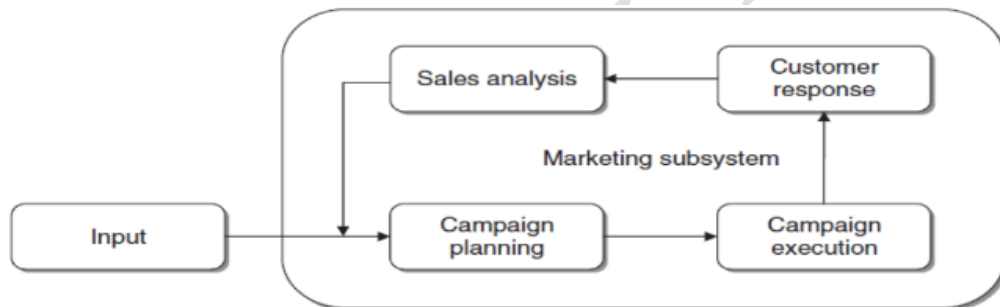
- The type of ethics in Business Intelligence (BI) is the ethical principles of conduct that govern an individual in the workplace or a company in general. It is also known as professional ethics and not to be confused with other forms of philosophical ethics including religious conviction, or popular conviction.
- Professional ethics according to Griffin (1986) is that profit is not the only important strategy of a business anymore. There is also more of a concern and motivator of companies to do what is right.
- Companies must acknowledge that they have a common good to protect their local community, improve employee relations and promote informational press to the public. While back in 1986, Griffin was directing his argument towards ethics in accounting but it is also true today in Business Intelligence.
- Government regulations are not changing fast enough to cover all the changes in technology that bombards users on day to day bases. It is up to corporations to create a code of ethics, and to persistently be receptive to the needs of the public being served. Everyday in BI management professionals may be at risk of making unethical practices in their decisions that regards the consumer, business and/or other

	<p>employees data. Ethics is a touchy subject, there is always going to be controversy on how companies choose to handle business decisions.</p> <ul style="list-style-type: none"> ➤ There is no definite decision to make when it comes to ethical decisions. While sometimes it may involve illegal practices, other times it is just a decision that needs to be made in a company to promote a better way of life for all. ➤ An example of an ethical decision would be a manager of a BI system that chooses to use cheaper data in his/her data mining activities to save money. The data he/she chooses to implement involves personal credit score reports. ➤ The cheaper data sets have a 20% possibility of being incorrect. The manager did not see it as being an unethical decision when it was made, just a way to continue to generate close-to-accurate reports and save money. ➤ The impacting decision on 20% of the company's customers may have different results as more people are turned down for credit because inaccurate reports. It is not a crime to have implemented the inaccurate data sets but it may seem as an unethical practice to others. ➤ While it is important for managers to be able to make their own decisions, this example decision being made should have involved more managers since it affected the whole business. ➤ The manager's choice could bankrupt the company as user start to leave their business for more accurate competitive companies. As the example points out, sometimes there is no really clear answer to wither an issue involves an ethical or legal choice and each situation can be different. ➤ Trying to make decisions based on individuals' beliefs when dealing with a company can amount to intellectual stalls and trying to come to a decision can be expensive and time consuming. ➤ Today's society has come to the point where there are more solutions to problems than ever before. What once was impossible can now be accomplished through the use of BI and other technology similar to BI.
6.	Describe the role of mathematical model.
	<ul style="list-style-type: none"> ➤ A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models. ➤ Mathematical models and algorithms help decision makers to extract information and knowledge from the data through the means of a business intelligence system. ➤ Data can be graphically represented by histograms, whereas more elaborate analysis requires development of advanced learning models. ➤ Generally, business intelligence system is used to promote a scientific and rational approach of organization. ➤ Example- a spreadsheet is used to estimate the effects on the fluctuations in interest rates with the help of that decision makers can generate a mental representation of the financial flows process. ➤ Classical scientific fields, such as physics, have always resorted to mathematical models for the abstract representation of real systems.

- Other areas, such as operations research, have instead made full use of the application of scientific methods and mathematical models to the study of artificial systems, for example public and private organizations.
- The rational approach typical of a business intelligence analysis can be summarized schematically in the following main characteristics.
 - First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
 - Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
 - Finally, what-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

7. Define system. Explain how system can be characterized.

System is made up of a set of components that are in some way connected to each other so as to provide a single collective result and a common purpose. Every system is characterized by boundaries that separate its internal components from the external environment.



Open Cycle System - A system is said to be open if its boundaries can be crossed in both directions by flows of materials and information. In general terms, any given system receives specific input flows, carries out an internal transformation process and generates observable output flows.

Closed Cycle System – Systems that are able to modify their own output flows based on feedback are called closed cycle systems.

For example, the closed cycle system outlined in the above figure. It describes the development of a sequence of marketing campaigns. The sales results for each campaign are gathered and become available as feedback input so as to design subsequent marketing promotions.

Abstract representation of a system :

A closed cycle marketing system with feedback effects

The sales results for each campaign are gathered and become available as feedback input so as to design subsequent marketing promotions.

In connection with a decision-making process, whose structure will be described in the next section, it is often necessary to assess the performance of a system. For this purpose, it is appropriate to categorize the evaluation metrics into two main classes: effectiveness and efficiency.

1. **Effectiveness**: Effectiveness measurements express the level of conformity of a given system to the objectives for which it was designed. The associated performance indicators are therefore linked to the system output flows, such as production volumes, weekly sales and yield per share.

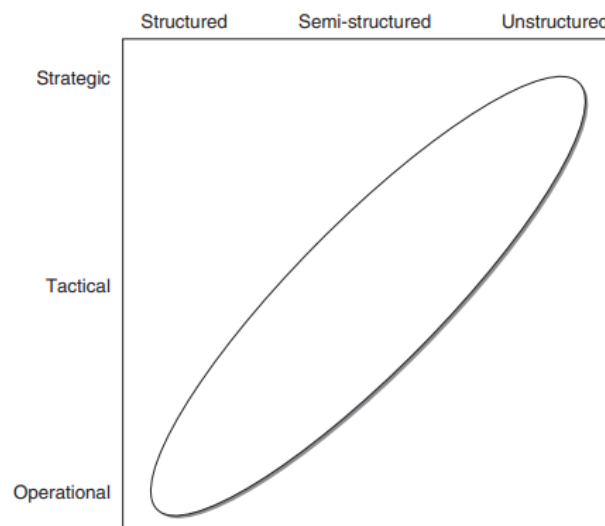
2. **Efficiency**: Efficiency measurements highlight the relationship between input flows used by

	the system and the corresponding output flows. Efficiency measurements are therefore associated with the quality of the transformation process. For example, they might express the amount of resources needed to achieve a given sales volume.
8.	What is decision support system (DSS)? What are the factors that affect the degree of success of the DSS?
	<p>A decision support system (DSS) is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations.</p> <p><u>The factors that affect the degree of success of the DSS are:</u></p> <p>Integration: The design and development of a DSS require a significant number of methodologies, tools, models, individuals and organizational processes to work in harmony. This results in a highly complex project requiring diverse competencies. The role of a system integrator is thus essential. A system integrator is an expert in all the aspects involved in the development of a DSS, such as information system architectures, decision-making processes, mathematical models and solution methods. This role is usually performed by a third party who may also exert a positive influence as an agent of innovation capable of overcoming most of the resistance to change that often arises in every organization.</p> <p>Involvement: The exclusion or marginalization from the project team of knowledge workers who will actually use the system once it is implemented is a mistake that is sometimes made during the design and development of a DSS. In many cases this happens because a DSS is mistakenly considered a mere computer application, the development of which is assigned solely or primarily to the information systems department.</p> <p>Uncertainty: In general, costs are not a major concern in the implementation of a DSS, and the advantage of devising more effective decisions largely offsets the development costs incurred. Of course, it is appropriate to reduce the project uncertainty through prototyping, user friendliness, system tests during the preliminary stages and an evolutionary implementation.</p>
9.	Explain classification of decisions according to their nature and scope.
	<p>According to their nature, decisions can be classified as structured, unstructured or semi-structured.</p> <p><u>Structured decisions:</u> A decision is structured if it is based on a well-defined and recurring decision-making procedure. In most cases structured decisions can be traced back to an algorithm, which may be more or less explicit for decision makers, and are therefore better suited for automation. More specifically, we have a structured decision if input flows, output flows and the transformations performed by the system can be clearly described in the three phases of intelligence, design and choice. In this case, we will also say that each component phase is structured in its turn.</p> <p><u>Unstructured decisions:</u> A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured. This means that for each phase there is at least one element in the system (input flows, output flows and the transformation processes) that cannot be described in detail and reduced to a predefined sequence of steps. Such an event may occur when a decision-making process is faced for the first time or if it happens very seldom. In this type of decisions the role of knowledge workers is fundamental, and business intelligence systems may provide support to decision makers through timely and versatile access to information.</p> <p><u>Semi-structured decisions:</u> A decision is semi-structured when some phases are structured and others are not. Most decisions faced by knowledge workers in managing public or private enterprises or organizations are semi-structured. Hence, they can take advantage of DSSs</p>

and a business intelligence environment primarily in two ways. For the unstructured phases of the decision-making process, business intelligence tools may offer a passive type of support which translates into timely and versatile access to information. For the structured phases it is possible to provide an active form of support through mathematical models and algorithms that allow significant parts of the decision-making process to be automated.

The nature of a decision process depends on many factors, including:

- the characteristics of the organization within which the system is placed;
- the subjective attitudes of the decision makers;
- the availability of appropriate problem-solving methodologies;
- the availability of effective decision support tools.



Depending on their **scope**, decisions can be classified as strategic, tactical and operational.

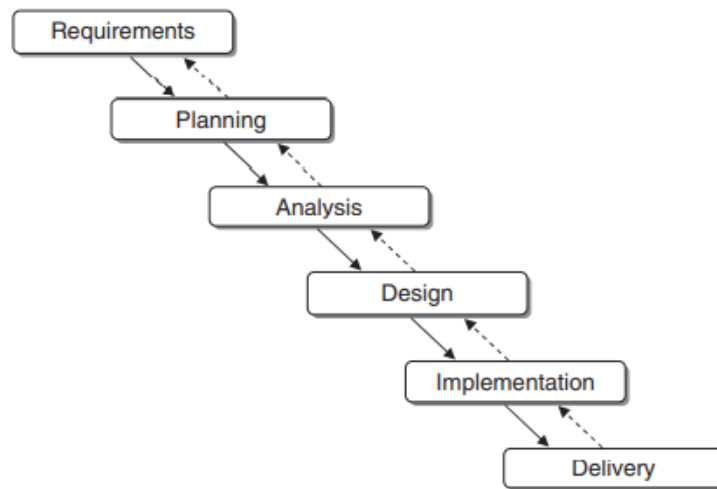
Strategic decisions: Decisions are strategic when they affect the entire organization or at least a substantial part of it for a long period of time. Strategic decisions strongly influence the general objectives and policies of an enterprise. As a consequence, strategic decisions are taken at a higher organizational level, usually by the company top management.

Tactical decisions: Tactical decisions affect only parts of an enterprise and are usually restricted to a single department. The time span is limited to a medium-term horizon, typically up to a year. Tactical decisions place themselves within the context determined by strategic decisions. In a company hierarchy, tactical decisions are made by middle managers, such as the heads of the company departments.

Operational decisions: Operational decisions refer to specific activities carried out within an organization and have a modest impact on the future. Operational decisions are framed within the elements and conditions determined by strategic and tactical decisions. Therefore, they are usually made at a lower organizational level, by knowledge workers responsible for a single activity or task such as sub-department heads, workshop foremen, back-office heads.

10. Describe different phases in the development of a decision support systems (DSS).

The major steps in the development of a DSS are shown below:



The logical flow of the activities is shown by the solid arrows. The dotted arrows in the opposite direction indicate revisions of one or more phases that might become necessary during the development of the system, through a feedback mechanism.

We describe now in detail how each phase is carried out.

Planning: The main purpose of the planning phase is to understand the needs and opportunities, sometimes characterized by weak signals, and to translate them into a project and later into a successful DSS. Planning usually involves a feasibility study to address the question: Why do we wish to develop a DSS? During the feasibility analysis, general and specific objectives of the system, recipients, possible benefits, execution times and costs are laid down.

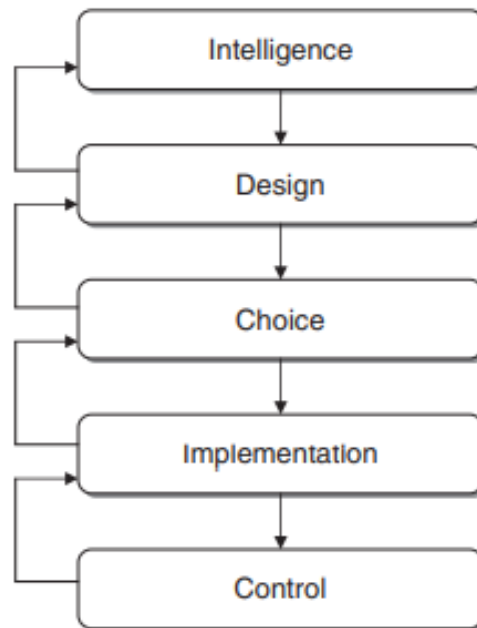
Analysis: In the analysis phase, it is necessary to define in detail the functions of the DSS to be developed, by further developing and elaborating the preliminary conclusions achieved during the feasibility study. A response should therefore be given to the following question: What should the DSS accomplish, and who will use it, when and how? To provide an answer, it is necessary to analyse the decision processes to be supported, to try to thoroughly understand all interrelations existing between the problems addressed and the surrounding environment. The analysis also involves mapping out the actual decision processes and imagining what the new processes will look like once the DSS is in place.

Design: During the design phase the main question is: How will the DSS work? The entire architecture of the system is therefore defined, through the identification of the hardware technology platforms, the network structure, the software tools to develop the applications and the specific database to be used. It is also necessary to define in detail the interactions with the users, by means of input masks, graphic visualizations on the screen and printed reports.

Implementation: Once the specifications have been laid down, it is time for implementation, testing and the actual installation, when the DSS is rolled out and put to work. Any problems faced in this last phase can be traced back to project management methods. A further aspect of the implementation phase, which is often overlooked, relates to the overall impact on the organization determined by the new system. Such effects should be monitored using change management techniques, making sure that no one feels excluded from the organizational innovation process and rejects the DSS.

11. Explain the phases of decision-making process system? Explain types of decision

A compelling representation of the decision-making process was proposed in the early 1960s, and still remains today a major methodological reference. The model includes three phases, termed intelligence, design and choice. An extended version of the original scheme, which results from the inclusion of two additional phases are namely implementation and control.



Intelligence: In the intelligence phase the task of the decision maker is to identify, circumscribe and explicitly define the problem that emerges in the system under study. The analysis of the context and all the available information may allow decision makers to quickly grasp the signals and symptoms pointing to a corrective action to improve the system performance. For example, during the execution of a project the intelligence phase may consist of a comparison between the current progress of the activities and the original development plan. In general, it is important not to confuse the problem with the symptoms.

Design: In the design phase actions aimed at solving the identified problem should be developed and planned. At this level, the experience and creativity of the decision makers play a critical role, as they are asked to devise viable solutions that ultimately allow the intended purpose to be achieved. Where the number of available actions is small, decision makers can make an explicit enumeration of the alternatives to identify the best solution. If, on the other hand, the number of alternatives is very large, or even unlimited, their identification occurs in an implicit way, usually through a description of the rules that feasible actions should satisfy. For example, these rules may directly translate into the constraints of an optimization model.

Choice: Once the alternative actions have been identified, it is necessary to evaluate them on the basis of the performance criteria deemed significant. Mathematical models and the corresponding solution methods usually play a valuable role during the choice phase. For example, optimization models and methods allow the best solution to be found in very complex situations involving countless or even infinite feasible solutions. On the other hand, decision trees can be used to handle decision-making processes influenced by stochastic events.

Implementation: When the best alternative has been selected by the decision maker, it is transformed into actions by means of an implementation plan. This involves assigning responsibilities and roles to all those involved into the action plan.

Control: Once the action has been implemented, it is finally necessary to verify and check that

	the original expectations have been satisfied and the effects of the action match the original intentions. In particular, the differences between the values of the performance indicators identified in the choice phase and the values actually observed at the end of the implementation plan should be measured. In an adequately planned DSS, the results of these evaluations translate into experience and information, which are then transferred into the data warehouse to be used during subsequent decision-making processes.
12.	Explain the major potential advantage derived from adoption of a DSS.
	<p>The major potential advantage derived from adoption of a DSS are:</p> <ul style="list-style-type: none"> ➤ an increase in the number of alternatives or options considered; ➤ an increase in the number of effective decisions devised; ➤ a greater awareness and a deeper understanding of the domain analysed and the problems investigated; ➤ the possibility of executing scenario and what-if analyses by varying the hypotheses and parameters of the mathematical models; ➤ an improved ability to react promptly to unexpected events and unforeseen situations; ➤ a value-added exploitation of the available data; ➤ an improved communication and coordination among the individuals and the organizational departments; ➤ more effective development of teamwork; ➤ a greater reliability of the control mechanisms, due to the increased intelligibility of the decision process.

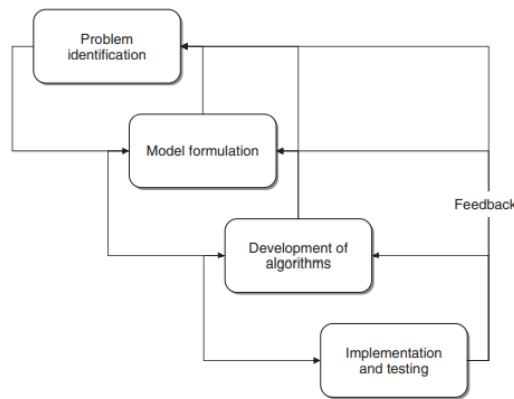
UNIT 2

1. Explain the divisions of mathematical models according to their characteristics, probabilistic nature, and temporal dimension.
2. What are the phases in the development of mathematical models for decision making? / Explain the primary phases of model.
3. Differentiate between supervised and unsupervised learning.
4. Explain classes of model. /What is predictive and optimization model?
5. What is data mining? List the real-life applications of data mining. / Explain some of the areas where data mining is used. / Write various application of data mining.
6. Write short notes on analysis methodology of data mining.
7. What is data mining process? / Draw and explain architecture of data mining.
8. Describe representation of input data.
9. Write short note on Corporate Analysis and risk management. And Fraud detection.
10. Explain categorical and numerical attributes with proper example.
11. Draw and explain data preparation./Write Short note on data preparation
12. What is meant by data validation? Explain different kinds of data validation.
13. Write short note on data transformation
14. Explain the following normalization techniques:
(i) Decimal scaling (ii) Min-max
15. Describe Data reduction.

1.	Explain the divisions of mathematical models according to their characteristics, probabilistic nature, and temporal dimension.
	<p><u>According to their characteristics, models can be divided into iconic, analogical and symbolic.</u></p> <p>Iconic: An iconic model is a material representation of a real system, whose behaviour is imitated for the purpose of the analysis. A miniaturized model of a new city neighbourhood is an example of iconic model.</p> <p>Analogical: An analogical model is also a material representation, although it imitates the real behaviour by analogy rather than by replication. A wind tunnel built to investigate the aerodynamic properties of a motor vehicle is an example of an analogical model intended to represent the actual progression of a vehicle on the road.</p> <p>Symbolic: A symbolic model, such as a mathematical model, is an abstract representation of a real system. It is intended to describe the behaviour of the system through a series of symbolic variables, numerical parameters and mathematical relationships.</p> <p><u>According to their probabilistic nature, models can be divided into either stochastic or deterministic.</u></p> <p>Stochastic: In a stochastic model some input information represents random events and is therefore characterized by a probability distribution, which in turn can be assigned or unknown. Predictive models, which will be thoroughly described in the following chapters, as well as waiting line models, briefly mentioned below in this chapter, are examples of stochastic models.</p> <p>Deterministic: A model is called deterministic when all input data are supposed to be known a priori and with certainty. Since this assumption is rarely fulfilled in real systems, one resorts to deterministic models when the problem at hand is sufficiently complex and any stochastic elements are of limited relevance. Notice, however, that even for deterministic models the hypothesis of knowing the data with certainty may be relaxed. Sensitivity and scenario analyses, as well as what-if analysis, allow one to assess the robustness of optimal decisions to variations in the input parameters.</p> <p><u>A further distinction concerns the temporal dimension in a mathematical model, which can be either static or dynamic.</u></p> <p>Static: Static models consider a given system and the related decision-making process within one single temporal stage.</p> <p>Dynamic: Dynamic models consider a given system through several temporal stages, corresponding to a sequence of decisions. In many instances the temporal dimension is subdivided into discrete intervals of a previously fixed span: minutes, hours, days, weeks, months and years are examples of discrete subdivisions of the time axis. Discrete-time dynamic models, which largely prevail in business intelligence applications, observe the status of a system only at the beginning or at the end of discrete intervals. Continuous-time dynamic models consider a continuous sequence of periods on the time axis.</p>

2. What are the phases in the development of mathematical models for decision making? / Explain the primary phases of model

It is possible to break down the development of a mathematical model for decision making into four primary phases.



Problem identification:

First of all, the problem at hand must be correctly identified. The observed critical symptoms must be analysed and interpreted in order to formulate hypothesis for investigation. For example, too high a stock level, corresponding to an excessive stock turnover rate, may possibly represent a symptom for a company manufacturing consumable goods. It is therefore necessary to understand what caused the problem, based on the opinion of the production managers. In this case, an ineffective production plan may be the cause of the stock accumulation.

Model formulation:

Once the problem to be analysed has been properly identified, effort should be directed toward defining an appropriate mathematical model to represent the system. A number of factors affect and influence the choice of model, such as the time horizon, the decision variables, the evaluation criteria, the numerical parameters and the mathematical relationships.

- Time horizon: Usually a model includes a temporal dimension. For example, to formulate a tactical production plan over the medium term it is necessary to specify the production rate for each week in a year, whereas to derive an operational schedule it is required to assign the tasks to each production line for each day of the week. As we can see, the time span considered in a model, as well as the length of the base intervals, may vary depending on the specific problem considered.
- Evaluation criteria: Appropriate measurable performance indicators should be defined in order to establish a criterion for the evaluation and comparison of the alternative decisions. These indicators may assume various forms in each different application, and may include the following factors:
 - monetary costs and payoffs;
 - effectiveness and level of service;
 - quality of products and services;
 - flexibility of the operating conditions;
 - reliability in achieving the objectives.
- Decision variables: Symbolic variables representing alternative decisions should then be defined. For example, if a problem consists of the formulation of a tactical production plan over the medium term, decision variables should express production volumes for each product, for each process and for each period of the planning horizon.

	<p>➤ Numerical parameters: It is also necessary to accurately identify and estimate all numerical parameters required by the model. In the production planning example, the available capacity should be known in advance for each process, as well as the capacity absorption coefficients for each combination of products and processes.</p> <p>➤ Mathematical relationships: The final step in the formulation of a model is the identification of mathematical relationships among the decision variables, the numerical parameters and the performance indicators defined during the previous phases. Sometimes these relationships may be exclusively deterministic, while in other instances it is necessary to introduce probabilistic relationships. In this phase, the trade-off between the accuracy of the representation achieved through the model and its solution complexity should be carefully considered. It may turn out more helpful at a practical level to adopt a model that sacrifices some marginal aspects of reality in the representation of the system but allows an efficient solution and greater flexibility in view of possible future developments.</p> <p>Development of algorithms: Once a mathematical model has been defined, one will naturally wish to proceed with its solution to assess decisions and to select the best alternative. In other words, a solution algorithm should be identified and a software tool that incorporates the solution method should be developed or acquired. An analyst in charge of model formulation should possess a thorough knowledge of current solution methods and their characteristics.</p> <p>Implementation and test: When a model is fully developed, then it is finally implemented, tested and utilized in the application domain. It is also necessary that the correctness of the data and the numerical parameters entered in the model be preliminarily assessed. These data usually come from a data warehouse or a data mart previously set up. Once the first numerical results have been obtained using the solution procedure devised, the model must be validated by submitting its conclusions to the opinion of decision makers and other experts in the application domain. A number of factors should be taken into account at this stage:</p> <ul style="list-style-type: none"> • the plausibility and likelihood of the conclusions achieved; • the consistency of the results at extreme values of the numerical parameters; • the stability of the results when minor changes in the input parameters are introduced.
3.	<p>Differentiate between supervised and unsupervised learning.</p> <p>Supervised learning: In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs. For example, in loyalty in the mobile phone industry, or expresses a measurable quantity, such as the total value of calls that will be placed by a customer in a future period. As a second example of the supervised perspective, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions. Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.</p> <p>Unsupervised learning: Unsupervised (or indirect) learning analyses are not guided by a target attribute. Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset. As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behaviour, based on data on past transactions. In most unsupervised learning analyses, one is interested in identifying clusters of records that are similar within each cluster and different from members of other clusters.</p>

4.	Explain classes of model. / What is predictive and optimization model?
	<p><u>Risk analysis model</u> Risk analysis is the process of assessing the likelihood of an adverse event occurring within the corporate, government, or environmental sector.</p> <p>Risk analysis is the study of the underlying uncertainty of a given course of action and refers to the uncertainty of forecasted cash flow streams, variance of portfolio/stock returns, the probability of a project's success or failure, and possible future economic states.</p> <p>Risk analysts often work in tandem with forecasting professionals to minimize future negative unforeseen effects.</p> <p><u>Project management model</u> Every project is extremely unique which means we cannot have a standard structure execute our projects and achieve success in our endeavor. However, to have a good plan we need some kind of framework or structure to follow depending on the nature of the project. Project management models or methodologies provide the framework to execute projects A framework is something that tells you how often you will meet and discuss the progress, how you will document results, how you will communicate and so on.</p> <p><u>Predictive models:</u> Predictive modelling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software. As additional data becomes available, the statistical analysis model is validated or revised.</p> <p><u>Optimization models:</u> As additional data becomes available, the statistical analysis model is validated or revised.</p> <p><u>Types of Optimization Models:</u> Optimization problems can be classified in terms of the nature of the objective function and the nature of the constraints. Special forms of the objective function and the constraints give rise to specialized algorithms that are more efficient. From this point of view, there are four types of optimization problems, of increasing complexity. An Unconstrained optimization problem is an optimization problem where the objective function can be of any kind (linear or nonlinear) and there are no constraints. These types of problems are handled by the classes discussed in the earlier sections.</p> <p>A linear program is an optimization problem with an objective function that is linear in the variables, and all constraints are also linear. Linear programs are implemented by the Linear Program class.</p> <p>A quadratic program is an optimization problem with an objective function that is quadratic in the variables (i.e. it may contain squares and cross products of the decision variables), and all constraints are linear. A quadratic program with no squares or cross products in the objective function is a linear program. Quadratic programs are implemented by, the Quadratic Program class.</p> <p>A nonlinear program is an optimization problem with an objective function that is an arbitrary nonlinear function of the decision variables, and the constraints can be linear or nonlinear. Nonlinear programs are implemented by the Nonlinear Program class.</p>

	<p><u>Waiting Line model</u></p> <p>There are basically two costs that must be balanced in waiting line system - the cost of service and the cost of waiting. Note that I am not considering another possible cost component - the cost of a scheduling system.</p> <p>Theoretically, a scheduling system is a management strategy designed to avoid waiting lines (meaning you should never wait in the doctor's office - yeah, right!) and is not covered in this module.</p> <p>Scheduling systems are useful when the customer is known to the system and the short and long run costs of waiting are relatively high. We will study scheduling system applications in linear programming later on in the course.</p> <p>Operational characteristics of waiting lines include:</p> <ol style="list-style-type: none"> 1. The probability that no customers (or units) are in the system. 2. The average number of customers in the lines. 3. The average number of customers in the system (customers in line plus those being served). 4. The average time a customer spends in the waiting line. 5. The average time a customer spends in the system (waiting time plus time in the service facility). 6. The probability that an arriving customer has to wait for service. <p><u>Pattern recognition model</u></p> <p>Pattern recognition deals with identifying a pattern and confirming it again. In general, a pattern can be a fingerprint image, a handwritten cursive word, a human face, a speech signal, a bar code, or a web page on the Internet. The individual patterns are often grouped into various categories based on their properties. When the patterns of same properties are grouped together, the resultant group is also a pattern, which is often called a pattern class.</p> <p>Pattern recognition is the science for observing, distinguishing the patterns of interest, and making correct decisions about the patterns or pattern classes. Thus, a biometric system applies pattern recognition to identify and classify the individuals, by comparing it with the stored templates.</p>
5.	<p>What is data mining? List the real-life applications of data mining. / Explain some of the areas where data mining is used. / Write various application of data mining.</p>
	<p>Areas where data mining is used are:</p> <p>Relational marketing: Data mining applications in the field of relational marketing have significantly contributed to the increase in the popularity of these methodologies. Some relevant applications within relational marketing are:</p> <ul style="list-style-type: none"> • identification of customer segments that are most likely to respond to targeted marketing campaigns, such as cross-selling and up-selling; • identification of target customer segments for retention campaigns; • prediction of the rate of positive responses to marketing campaigns; • interpretation and understanding of the buying behaviour of the customers; • analysis of the products jointly purchased by customers, known as market basket analysis. <p>Fraud detection: Fraud detection is another relevant field of application of data mining. Fraud may affect different industries such as telephony, insurance (false claims) and banking (illegal use of credit cards and bank checks; illegal monetary transactions).</p> <p>Risk evaluation: The purpose of risk analysis is to estimate the risk connected with future decisions, which often assume a dichotomous form. For example, using the past observations available, a bank may develop a predictive model to establish if it is appropriate to grant a monetary loan or a home loan, based on the characteristics of the applicant.</p> <p>Text mining: Data mining can be applied to different kinds of texts, which represent</p>

	<p>unstructured data, in order to classify articles, books, documents, emails and web pages. Examples are web search engines or the automatic classification of press releases for storing purposes. Other text mining applications include the generation of filters for email messages and newsgroups.</p> <p>Image recognition: The treatment and classification of digital images, both static and dynamic, is an exciting subject both for its theoretical interest and the great number of applications it offers. It is useful to recognize written characters, compare and identify human faces, apply correction filters to photographic equipment and detect suspicious behaviours through surveillance video cameras.</p> <p>Web mining: Web mining applications are intended for the analysis of so-called clickstreams – the sequences of pages visited and the choices made by a web surfer. They may prove useful for the analysis of e-commerce sites, in offering flexible and customized pages to surfers, in caching the most popular pages or in evaluating the effectiveness of an e-learning training course.</p> <p>Medical diagnosis: Learning models are an invaluable tool within the medical field for the early detection of diseases using clinical test results. Image analysis for diagnostic purpose is another field of investigation that is currently burgeoning.</p>
6.	<p>Write short notes on analysis methodology of data mining.</p> <p>Data mining activities can be subdivided into a few major categories, based on the tasks and the objectives of the analysis. Depending on the possible existence of a target variable, one can draw a first fundamental distinction between supervised and unsupervised learning processes.</p> <p>Supervised learning: In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs, on loyalty in the mobile phone industry, or expresses a measurable quantity, such as the total value of calls that will be placed by a customer in a future period. As a second example of the supervised perspective, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions. Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.</p> <p>Unsupervised learning: Unsupervised (or indirect) learning analyses are not guided by a target attribute. Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset. As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behaviour, based on data on past transactions. In most unsupervised learning analyses, one is interested in identifying clusters of records that are similar within each cluster and different from members of other clusters. Taking the distinction even further, seven basic data mining tasks can be identified:</p> <ul style="list-style-type: none"> • characterization and discrimination; • classification; • regression; • time series analysis; • association rules; • clustering; • description and visualization. <p>The first four tasks correspond to supervised data mining analyses, since a specific target variable exists that must be explained based on the available attributes or throughout its evolution over time. The remaining three tasks represent unsupervised analyses whose purpose is the development of models capable of expressing the interrelationship among the available attributes.</p>

Characterization and discrimination:

Where a categorical target attribute exists, before starting to develop a classification model, it is often useful to carry out an exploratory analysis whose purpose is twofold. On the one hand, the aim is to achieve a characterization by comparing the distribution of the values of the attributes for the records belonging to the same class.

Classification:

In a classification problem a set of observations is available, usually represented by the records of a dataset, whose target class is known. Observations may correspond, for instance, to mobile phone customers and the binary class may indicate whether a given customer is still active or has churned. Each observation is described by a given number of attributes whose value is known; in the previous example, the attributes may correspond to age, customer seniority and outgoing telephone traffic distinguished by destination.

Regression:

Unlike classification, which is intended for discrete targets, regression is used when the target variable takes on continuous values. Based on the available explanatory attributes, the goal is to predict the value of the target variable for each observation. If one wishes to predict the sales of a product based on the promotional campaigns mounted and the sale price, the target variable may take on a very high number of discrete values and can be treated as a continuous variable

Time series:

Sometimes the target attribute evolves over time and is therefore associated with adjacent periods on the time axis. In this case, the sequence of values of the target variable is said to represent a time series. For instance, the weekly sales of a given product observed over 2 years represent a time series containing 104 observations. Models for time series analysis investigate data characterized by a temporal dynamics and are aimed at predicting the value of the target variable for one or more future periods.

Association rules:

Association rules, also known as affinity groupings, are used to identify interesting and recurring associations between groups of records of a dataset. For example, it is possible to determine which products are purchased together in a single transaction and how frequently. Companies in the retail industry resort to association rules to design the arrangement of products on shelves or in catalogues. Groupings by related elements are also used to promote cross-selling or to devise and promote combinations of products and services.

Clustering: The term cluster refers to a homogeneous subgroup existing within a population. Clustering techniques are therefore aimed at segmenting a heterogeneous population into a given number of subgroups composed of observations that share similar characteristics; observations included in different clusters have distinctive features. Unlike classification, in clustering there are no predefined classes or reference examples indicating the target class, so that the objects are grouped together based on their mutual homogeneity

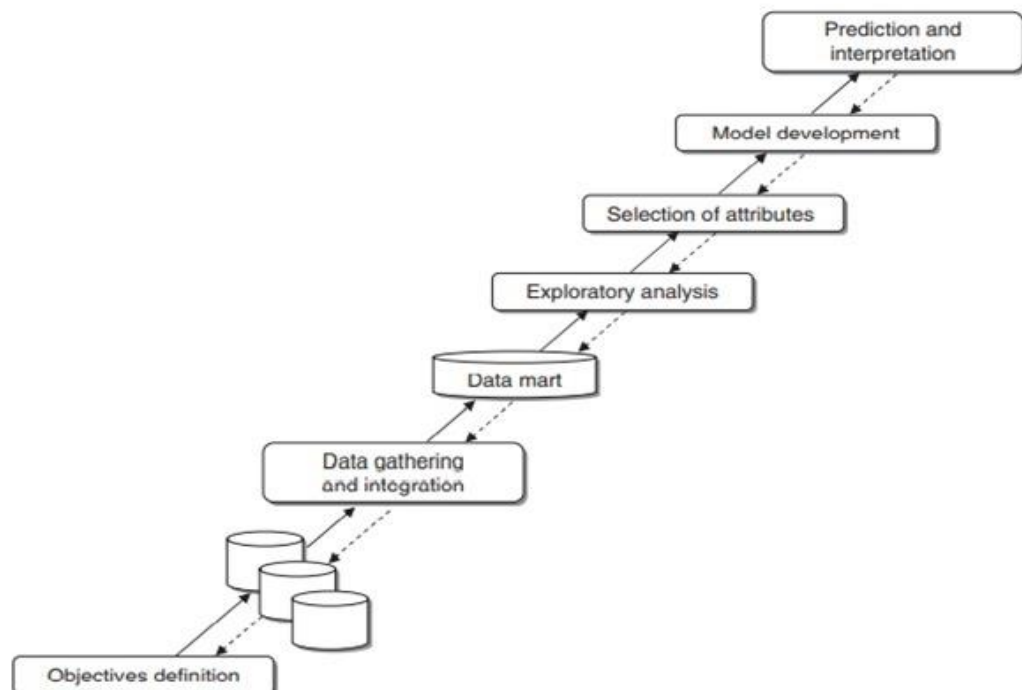
Description and visualization: The purpose of a data mining process is sometimes to provide a simple and concise representation of the information stored in a large dataset. Although, in contrast to clustering and association rules, the descriptive analysis does not pursue any particular grouping or partition of the records in the dataset, an effective and concise description of information is very helpful, since it may suggest possible explanations of hidden patterns in the data and lead to a better understanding the phenomena to which the data refer.

7. What is data mining process? / Draw and explain architecture of data mining.

The definition of data mining given at the beginning of an iterative process, during which learning models and techniques play a key, though non-exhaustive, role. The main phases of a generic data mining process.

Definition of objectives:

Data mining analyses are carried out in specific application domains and are intended to provide decision makers with useful knowledge. As a consequence, intuition and competence are required by the domain experts in order to formulate plausible and well-defined investigation objectives. If the problem at hand is not adequately identified and circumscribed one may run the risk of thwarting any future effort made during data mining activities. The definition of the goals will benefit from close cooperation between experts in the field of application and data mining analysts. It is possible to define the problem and the goals of the investigation as the analysis of past data and identification of a model so as to express the propensity of customers to leave the service (churn) based on their characteristics, in order to understand the reasons for such disloyalty and predict the probability of churn.



Data gathering and integration:

Once the objectives of the investigation have been identified, the gathering of data begins. Data may come from different sources and therefore may require integration. Data sources may be internal, external or a combination of the two. The integration of distinct data sources may be suggested by the need to enrich the data with new descriptive dimensions, such as geomarketing variables, or with lists of names of potential customers, termed prospects, not yet existing in the company information system.

Data mart:

In some instances, data sources are already structured in data warehouses and data marts for OLAP analyses and more generally for decision support activities. These are favorable situations where it is sufficient to select the attributes deemed relevant for the purpose of a data mining analysis. There is a risk, however, that, in order to limit memory uptake, the information stored in a data warehouse has been aggregated and consolidated to such an extent as to render useless any subsequent analysis.

Exploratory analysis:

In the third phase of the data mining process, a preliminary analysis of the data is carried out with the purpose of getting acquainted with the available information and carrying out data cleansing. Usually, the data stored in a data warehouse are processed at loading time in such a way as to remove any syntactical inconsistencies. For example, dates of birth that fall outside admissible ranges and negative sales charges are detected and corrected. In the data mining process, data cleansing occurs at a semantic level. First of all, the distribution of the values for each attribute is studied, using histograms for categorical attributes and basic summary statistics for numerical variables. In this way, any abnormal values (outliers) and missing values are also highlighted.

Attribute Selection:

In the subsequent phase, the relevance of the different attributes is evaluated in relation to the goals of the analysis. Attributes that prove to be of little use are removed, in order to cleanse irrelevant information from the dataset. Furthermore, new attributes obtained from the original variables through appropriate transformations are included into the dataset. For example, in most cases it is helpful to introduce new attributes that reflect the trends inherent in the data through the calculation of ratios and differences between original variables. Exploratory analysis and attribute selection are critical and often challenging stages of the data mining process and may influence to a great extent the level of success of the subsequent stages.

Model development and validation:

Once a high quality dataset has been assembled and possibly enriched with newly defined attributes, pattern recognition and predictive models can be developed. Usually the training of the models is carried out using a sample of records extracted from the original dataset. Then, the predictive accuracy of each model generated can be assessed using the rest of the data. More precisely, the available dataset is split into two subsets. The first constitutes the training set and is used to identify a specific learning model within the selected class of models

Prediction and interpretation:

Upon conclusion of the data mining process, the model selected among those generated during the development phase should be implemented and used to achieve the goals that were originally identified. Moreover, it should be incorporated into the procedures supporting decision-making processes so that knowledge workers may be able to use it to draw predictions and acquire a more in-depth knowledge of the phenomenon of interest.

8. Describe representation of input data.

In most cases, the input to a data mining analysis takes the form of a two dimensional table, called a dataset, irrespective of the actual logic and material representation adopted to store the information in files, databases, data warehouses and data marts used as data sources. The rows in the dataset correspond to the observations recorded in the past and are also called examples, cases, instances or records. The columns represent the information available for each observation and are termed attributes, variables, characteristics or features. The attributes contained in a dataset can be categorized as categorical or numerical, depending on the type of values they take on.

Categorical:

Categorical attributes assume a finite number of distinct values, in most cases limited to less than a hundred, representing a qualitative property of an entity to which they refer. Examples of categorical attributes are the province of residence of an individual (which takes as values a series of names, which in turn may be represented by integers) or whether a customer has abandoned her service provider (expressed by the value 1) or remained loyal to it (expressed by the value 0). Arithmetic operations cannot be applied to categorical attributes even when the coding of their values is expressed by integer numbers.

Numerical:

Numerical attributes assume a finite or infinite number of values and lend themselves to

subtraction or division operations. For example, the amount of outgoing phone calls during a month for a generic customer represents a numerical variable. Regarding two customers A and B making phone calls in a week for €27 and €36 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to €9 and that A has spent three fourths of the amount spent by B. Sometimes a more refined taxonomy of attributes can prove useful.

Counts:

Counts are categorical attributes in relation to which a specific property can be true or false. These attributes can therefore be represented using Boolean variables {true, false} or binary variables {0,1}. For example, a bank's customers may or may not be holders of a credit card issued by the bank. Nominal. Nominal attributes are categorical attributes without a natural ordering, such as the province of residence. Ordinal. Ordinal attributes, such as education level, are categorical attributes that lend themselves to a natural ordering but for which it makes no sense to calculate differences or ratios between the values. Discrete. Discrete attributes are numerical attributes that assume a finite number or a countable infinity of values.

9. Write short note on Corporate Analysis and risk management. And Fraud detection.

Data mining is used in the following fields of the Corporate Sector:

Finance Planning and Asset Evaluation: It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.

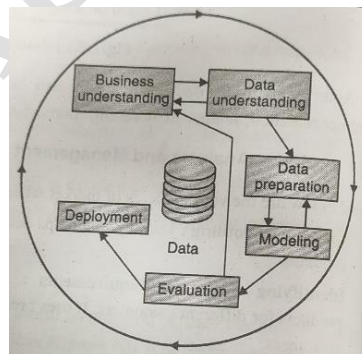
Resource Planning: It involves summarizing and comparing the resources and spending.

Competition: It involves monitoring competitors and market directions.

Fraud Detection:

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 analyses the patterns that deviate from expected norms.

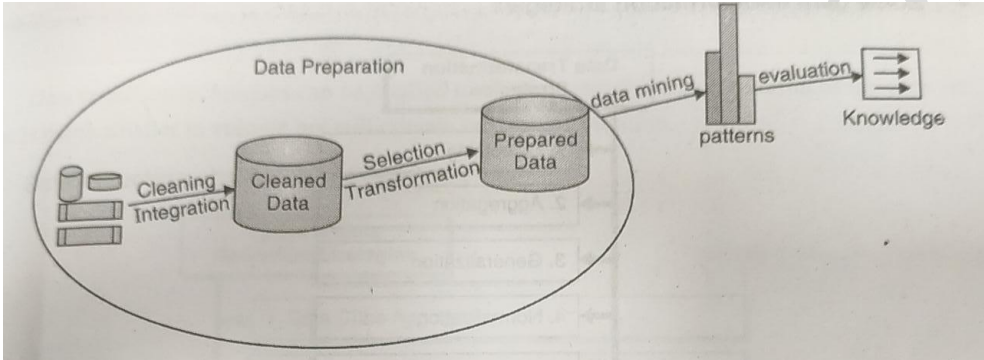


1. Business understanding In the business understanding phase:

First, it is required to understand business objectives clearly and find out what are the business's needs. Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered. Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation.

Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible. First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data.

	<p>2. Data understanding Some important activities must be performed including data load and data integration in order to make the data collection successfully. Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported. Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization. Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"</p> <p>3. Data preparation The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.</p> <p>4. Modeling First, modeling techniques have to be selected to be used for the prepared dataset. Next, the test scenario must be generated to validate the quality and validity of the model.</p>
10.	Explain categorical and numerical attributes with proper example.
	<p>The attributes contained in a dataset can be categorized as categorical or numerical, depending on the type of values they take on.</p> <p>Categorical: Categorical attributes assume a finite number of distinct values, in most cases limited to less than a hundred, representing a qualitative property of an entity to which they refer. Examples of categorical attributes are the province of residence of an individual (which takes as values a series of names, which in turn may be represented by integers) or whether a customer has abandoned her service provider (expressed by the value 1) or remained loyal to it (expressed by the value 0). Arithmetic operations cannot be applied to categorical attributes even when the coding of their values is expressed by integer numbers.</p> <p>Numerical: Numerical attributes assume a finite or infinite number of values and lend themselves to subtraction or division operations. For example, the amount of outgoing phone calls during a month for a generic customer represents a numerical variable. Regarding two customers A and B making phone calls in a week for €27 and €36 respectively, it makes sense to claim that the difference between the amounts spent by the two customers is equal to €9 and that A has spent three fourths of the amount spent by B.</p>

11.	Draw and explain data preparation./Write Short note on data preparation
	<p>Data preparation (or data pre-processing) in this context means manipulation of data into a form suitable for further analysis and processing. It is a process that involves many different tasks and which cannot be fully automated.</p> <p>Many of the data preparation activities are routine, tedious, and time consuming. It has been estimated that data preparation accounts for 60%-80% of the time spent on a data mining project.</p> <p>Data preparation is essential for successful data mining. Poor quality data typically result in incorrect and unreliable data mining results.</p> <p>Data preparation improves the quality of data and consequently helps improve the quality of data mining results. The well-known saying "garbage-in garbage-out" is very relevant to this domain.</p> 
12.	What is meant by data validation? Explain different kinds of data validation.
	<p>Data validation is about checking the information and to ensure that it complements the data needs of the system. This removes the chances of errors. One of the many examples of data validation is range check.</p> <p>Data validation has nothing to do with what the user wants to input. Validation is about checking the input data to ensure it conforms to the data requirements of the system to avoid data errors.</p> <p>An example of this is a range check to avoid an input number that is greater or smaller than the specified range.</p> <p>The quality of input data may prove unsatisfactory due to incompleteness, noise and inconsistency.</p> <p>Incompleteness – Some records may contain missing values corresponding to one or more attributes, and there may be a variety of reasons for this. It may be that some data were not recorded at the source in a systematic way, or that they were not available when the transactions associated with a record took place. In other instances, data may be missing because of malfunctioning recording devices. It is also possible that some data were deliberately removed during previous stages of the gathering process because they were deemed incorrect. Incompleteness may also derive from a failure to transfer data from the operational databases to a data mart used for a specific business intelligence analysis.</p> <p>Noise – Data may contain erroneous or anomalous values, which are usually referred to as outliers. Other possible causes of noise are to be sought in malfunctioning devices for data measurement, recording and transmission. The presence of data expressed in heterogeneous measurement units, which therefore require conversion, may in turn cause anomalies and inaccuracies.</p> <p>Inconsistency – Sometimes data contain discrepancies due to changes in the coding system</p>

	used for their representation, and therefore may appear inconsistent. For example, the coding of the products manufactured by a company may be subject to a revision taking effect on a given date, without the data recorded in previous periods being subject to the necessary transformations in order to adapt them to the revised encoding scheme. The purpose of data validation techniques is to identify and implement corrective actions in case of incomplete and inconsistent data or data affected by noise.
13.	<p>Write short note on data transformation</p> <p>In most data mining analyses it is appropriate to apply a few transformations to the dataset in order to improve the accuracy of the learning models subsequently developed. Outlier correction techniques are examples of transformations of the original data that facilitate subsequent learning phases.</p> <ul style="list-style-type: none"> Standardization: Most learning models benefit from a preventive standardization of the data, also called normalization. The most popular standardization techniques include the decimal scaling method, the min-max method and the z-index method. <ol style="list-style-type: none"> Decimal scaling: Decimal scaling is based on the transformation $x'_{ij} = \frac{x_{ij}}{10^h},$ where h is a given parameter which determines the scaling intensity. In practice, decimal scaling corresponds to shifting the decimal point by h positions toward the left. In general, h is fixed at a value that gives transformed values in the range [-1, 1]. Min-max: Min-max standardization is achieved through the transformation $x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}} (x'_{\max,j} - x'_{\min,j}) + x'_{\min,j},$ Where $x_{\min,j} = \min_i x_{ij}, \quad x_{\max,j} = \max_i x_{ij},$ are the minimum and maximum values of the attribute a_j before transformation, while $x'_{\min,j}$ and $x'_{\max,j}$ are the minimum and maximum values that we wish to obtain after transformation. In general, the extreme values of the range are defined so that $x'_{\min,j} = -1$ and $x'_{\max,j} = 1$ or $x'_{\min,j} = 0$ and $x'_{\max,j} = 1$. z-index: z-index based standardization uses the transformation $x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$ where $\bar{\mu}_j$ and $\bar{\sigma}_j$ are respectively the sample mean and sample standard deviation of the attribute a_j. If the distribution of values of the attribute a_j is roughly normal, the z-index based transformation generates values that are almost certainly within the range (-3, 3). <ul style="list-style-type: none"> Feature extraction: The aim of standardization techniques is to replace the values of an attribute with values obtained through an appropriate transformation. However, there are situations in which more complex transformations are used to generate new attributes that represent a set of additional columns in the matrix X representing the dataset D. Transformations of this kind are usually referred to as feature extraction. In some instances, the transformations may take even more complex forms, such as Fourier transforms, wavelets and kernel functions. The use of such methods will be explained within the classification methods called support vector machines. Attribute extraction may also consist of the creation of new variables that summarize within themselves the relevant information contained in a subset of the original attributes.

14.	<p>Explain the following normalization techniques: (i) Decimal scaling (ii) Min-max</p>
	<p>Decimal scaling: Decimal scaling is based on the transformation</p> $x'_{ij} = \frac{x_{ij}}{10^h},$ <p>where h is a given parameter which determines the scaling intensity. In practice, decimal scaling corresponds to shifting the decimal point by h positions toward the left. In general, h is fixed at a value that gives transformed values in the range [-1, 1].</p> <p>Min-max: Min-max standardization is achieved through the transformation</p> $x'_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}} (x'_{\max,j} - x'_{\min,j}) + x'_{\min,j},$ <p>Where</p> $x_{\min,j} = \min_i x_{ij}, \quad x_{\max,j} = \max_i x_{ij},$ <p>are the minimum and maximum values of the attribute a_j before transformation, while $x_{\min,j}$ and $x_{\max,j}$ are the minimum and maximum values that we wish to obtain after transformation. In general, the extreme values of the range are defined so that $x_{\min,j} = -1$ and $x_{\max,j} = 1$ or $x_{\min,j} = 0$ and $x_{\max,j} = 1$.</p>
15.	<p>Describe Data reduction.</p>
	<p>When dealing with a small dataset, the transformations are usually adequate to prepare input data for a data mining analysis. However, when facing a large dataset it is also appropriate to reduce its size, in order to make learning algorithms more efficient, without sacrificing the quality of the results obtained.</p> <p>There are three main criteria to determine whether a data reduction technique should be used: efficiency, accuracy and simplicity of the models generated.</p> <p>Efficiency: The application of learning algorithms to a dataset smaller than the original one usually means a shorter computation time. If the complexity of the algorithm is a super linear function, as is the case for most known methods, the improvement in efficiency resulting from a reduction in the dataset size may be dramatic. Within the data mining process it is customary to run several alternative learning algorithms in order to identify the most accurate model. Therefore, a reduction in processing times allows the analyses to be carried out more quickly.</p> <p>Accuracy: In most applications, the accuracy of the models generated represents a critical success factor, and it is therefore the main criterion followed in order to select one class of learning methods over another. As a consequence, data reduction techniques should not significantly compromise the accuracy of the model generated. It may also be the case that some data reduction techniques, based on attribute selection, will lead to models with a higher generalization capability on future records.</p> <p>Simplicity: In some data mining applications, concerned more with interpretation than with prediction, it is important that the models generated be easily translated into simple rules that can be understood by experts in the application domain. As a trade-off for achieving simpler rules, decision makers are sometimes willing to allow a slight decrease in accuracy. Data reduction often represents an effective technique for deriving models that are more easily interpretable.</p> <p>Data reduction can be pursued in three distinct directions, described below: a reduction in the number of observations through sampling, a reduction in the number of attributes through selection and projection, and a reduction in the number of values through discretization and aggregation.</p>

UNIT 3

1. Explain classification problems in detail /Explain Phases and Taxonomy of classification model.
2. Write short notes on evaluation of classification model. OR What are the criteria used to evaluate classification methods?
3. Write a short note on confusion matrix.
4. Explain top-down induction of decision tree. Examine the components of the top-down induction of decision trees procedure
5. Write a short note on Bayesian methods.? Assume your own training dataset and predict the class label of an unknown sampling using Naïve Bayesian classification.
6. Describe Logistic regression.
7. Describe neural networks/Explain the 'Rosenblatt perceptron' form of neural network with diagram
8. Write short note on: Clustering methods/What is taxonomy of clustering method? / What are the characteristics of clustering method?
9. What is clustering attributes? Write a short note on:
 - A) Binary Attribute
 - B) Nominal Attribute
 - C) Ordinal Attribute
10. Differentiate between following clustering methodologies:
 - i) Partitioning method.
 - ii) Hierarchical method.
11. Write k-means algorithm for clustering. / Explain k-means method.
12. Explain evaluation of clustering model.

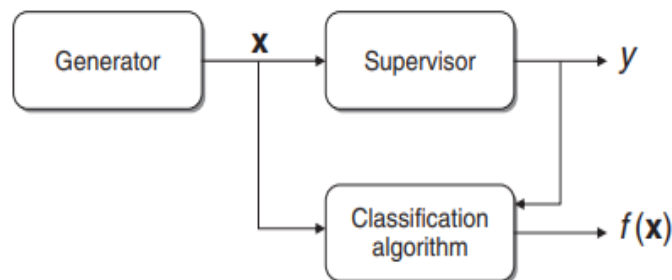
1. Explain classification problems in detail / Explain Phases and Taxonomy of classification model.

Classification Problems:

In a classification problem, we have a dataset D containing m observations described in terms of n explanatory attributes and a categorical target attribute.

The explanatory attributes, also called predictive variables, may be partly categorical and partly numerical.

The target attribute is also called a class or label, while the observations are also termed examples or instances. Unlike regression, for classification models the target variable takes a finite number of values. In particular, we have a binary classification problem if the instances belong to two classes only, and a multiclass or multcategory classification if there are more than two classes.



Probabilistic structure of the learning process for classification

The purpose of a classification model is to identify recurring relationships among the explanatory variables which describe the examples belonging to the same class. Such relationships are then translated into classification rules which are used to predict the class of examples for which only the values of the explanatory attributes are known. The rules may take different forms depending on the type of model used.

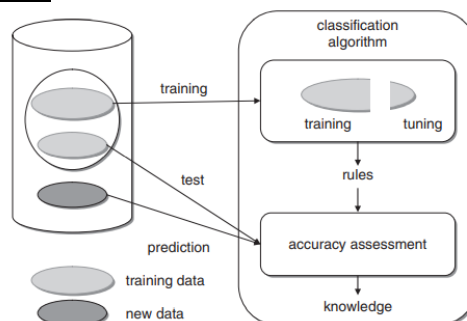
There are three component of classification problems. They are : a generator of observations, a supervisor of the target class and a classification algorithm.

Generator: The task of the generator is to extract random vectors x of examples according to an unknown probability distribution $P_x(x)$.

Supervisor: The supervisor returns for each vector x of examples the value of the target class according to a conditional distribution $P_{y|x}(y|x)$ which is also unknown.

Algorithm: A classification algorithm A_F , also called a classifier, chooses a function $f^* \in F$ in the hypothesis space so as to minimize a suitably defined loss function.

Phases of Classification models:



Phases of the learning process for a classification algorithm

The three main phases of classification model are as follows:

1. Training phase:

The classification algorithm is applied to the subnet of N which is called as training set.

To derive classification rules it allow the corresponding target class z to be involved to each observation m .

2. Test phase:

The rules are generated during the training phase. It is used to classify the observations of N . It is not included in the training set, for which the target class value is already known. The training set and test set should be different.

3. Prediction phase:

A prediction is achieved by applying the rules generated during the training phase to the explanatory variables that describe the new instance.

Taxonomy of classification models:

There are four main categories of classification models.

Heuristic models:

Heuristic methods make use of classification procedures based on simple and intuitive algorithms. This category includes nearest neighbour methods, based on the concept of distance between observations, and classification trees, which use divide-and-conquer schemes to derive groups of observations that are as homogeneous as possible with respect to target class.

Separation models:

Separation models divide the attribute space R^n into H disjoint regions $\{S_1, S_2, \dots, S_H\}$, separating the observations based on the target class. The observations x_i in region S_h are assigned to class $y_i = v_h$. Each region S_h may comprise a composite set obtained from set-theoretic operations of union and intersection applied to regions of an elementary form, such as half-spaces or hyperspheres. However, the resulting regions should be kept structurally simple, to avoid compromising the generalization capability of the model. The various classification models that belong to this category differ from one another in terms of the type of separation regions, the loss function, the algorithm used to solve the optimization problem. The most popular separation techniques include discriminant analysis, perceptron methods, neural networks and support vector machines. Some variants of classification trees can also be placed in this category.

Regression models:

Regression models, for the prediction of continuous target variables, make an explicit assumption concerning the functional form of the conditional probabilities $P_{y|x}(y|x)$, which correspond to the assignment of the target class by the supervisor. For linear regression models, it is assumed that a linear relationship exists between the dependent variable and the predictors, and this leads to the derivation of the value of the regression coefficients.

Probabilistic models:

In probabilistic models, a hypothesis is formulated regarding the functional form of the conditional probabilities $P_{x|y}(x|y)$ of the observations given the target class, known as class-conditional probabilities. Subsequently, based on an estimate of the prior probabilities $P_y(y)$ and using Bayes' theorem, the posterior probabilities $P_{y|x}(y|x)$ of the target class assigned by the supervisor can be calculated. The functional form of $P_{x|y}(x|y)$ may be parametric or nonparametric. Naive Bayes classifiers and Bayesian networks are well-known families of probabilistic methods.

2.	<p>Write short notes on evaluation of classification model. OR What are the criteria used to evaluate classification methods?</p>
	<p><u>Evaluation of classification model:</u> Within a classification analysis it is usually advisable to develop alternative models and then select the method affording the best prediction accuracy. To obtain alternative models one may vary the method, by using for instance classification trees, neural networks, Bayesian techniques or support vector machines, and also modify the values of the parameters involved. Classification methods can be evaluated based on several criteria, as follows.</p> <p>Accuracy: Evaluating the accuracy of a classification model is crucial for two main reasons. First, the accuracy of a model is an indicator of its ability to predict the target class for future observations. Based on their accuracy values, it is also possible to compare different models in order to select the classifier associated with the best performance. Let T be the training set and V the test set, and t and v be the number of observations in each subset, respectively. The relations $D = T \cup V$ and $m = t + v$ obviously hold. The most natural indicator of the accuracy of a classification model is the proportion of observations of the test set V correctly classified by the model. If y_i denotes the class of the generic observation $x_i \in V$ and $f(x_i)$ the class predicted through the function $f \in F$ identified by the learning algorithm $A = AF$, the following loss function can be defined:</p> $L(y_i, f(x_i)) = \begin{cases} 0, & \text{if } y_i = f(x_i), \\ 1, & \text{if } y_i \neq f(x_i). \end{cases}$ <p>The accuracy of model A can be evaluated as</p> $\text{acc}_A(V) = \text{acc}_{AF}(V) = 1 - \frac{1}{v} \sum_{i=1}^v L(y_i, f(x_i)).$ <p>In some cases, it is preferable to use an alternative performance indicator given by the proportion of errors made by the classification algorithm:</p> $\text{err}_A(V) = \text{err}_{AF}(V) = 1 - \text{acc}_{AF}(V) = \frac{1}{v} \sum_{i=1}^v L(y_i, f(x_i)).$ <p>Speed: Some methods require shorter computation times than others and can handle larger problems. However, classification methods characterized by longer computation times may be applied to a small-size training set obtained from a large number of observations by means of random sampling schemes. It is not uncommon to obtain more accurate classification rules in this way.</p> <p>Robustness: A classification method is robust if the classification rules generated, as well as the corresponding accuracy, do not vary significantly as the choice of the training set and the test set varies, and if it is able to handle missing data and outliers.</p> <p>Scalability: The scalability of a classifier refers to its ability to learn from large datasets, and it is inevitably related to its computation speed. Therefore, the remarks made in connection with sampling techniques for data reduction, which often result in rules having better generalization capability, also apply in this case.</p> <p>Interpretability: If the aim of a classification analysis is to interpret as well as predict, then the rules generated should be simple and easily understood by knowledge workers and experts in the application domain.</p>

3. Write a short note on confusion matrix.

The accuracy measurement methods described above are not always adequate for discriminating among models, and in some instances they may even yield paradoxical results it is useful to resort to decision tables, usually called confusion matrices, which for the sake of simplicity we will only describe in connection with binary classification, though they can be easily extended to multiclass classification. Let us assume that we wish to analyze a binary classification problem where the values taken by the target class are $\{-1, 1\}$. We can then consider a 2×2 matrix whose rows correspond to the observed values and whose columns are associated with the values predicted using a classification model. The elements of the confusion matrix have the following meanings: p is the number of correct predictions for the negative examples, called true negatives; u is the number of incorrect predictions for the positive examples, called false negatives; q is the number of incorrect predictions for the negative examples, called false positives; and v is the number of correct predictions for the positive examples, called true positives. Using these elements, further indicators useful for validating a classification algorithm can be defined

		predictions		
		-1 (negative)	+1 (positive)	total
examples	-1 (negative)	p	q	$p + q$
	+1 (positive)	u	v	$u + v$
	total	$p + u$	$q + v$	m

Accuracy. The accuracy of a classifier may be expressed as

$$\text{acc} = \frac{p + v}{p + q + u + v} = \frac{p + v}{m}$$

True negatives rate. The true negatives rate is defined as

$$\text{tn} = \frac{p}{p + q}$$

False negatives rate. The false negatives rate is defined as

$$\text{fn} = \frac{u}{u + v}$$

False positives rate. The false positives rate is defined as

$$\text{fp} = \frac{q}{p + q}$$

True positives rate. The true positives rate, also known as recall, is defined as

$$\text{tp} = \frac{v}{u + v}$$

Precision. The precision is the proportion of correctly classified positive examples, and is given by

$$\text{prc} = \frac{v}{q + v}$$

Geometric mean. The geometric mean is defined as

$$\text{gm} = \sqrt{\text{tp} \times \text{prc}}.$$

and sometimes also as

$$\text{gm} = \sqrt{\text{tp} \times \text{tn}}.$$

F-measure. The F-measure is defined as

$$F = \frac{(\beta^2 - 1) \text{tp} \times \text{prc}}{\beta^2 \text{prc} + \text{tp}}$$

where $\beta \in [0, \infty)$ regulates the relative importance of the precision with respect to the true positives rate. The F-measure is also equal to 0 if all the predictions are incorrect

4. Explain top-down induction of decision tree. Examine the components of the top-down induction of decision trees procedure

Top-down induction of decision tree

Classification trees are perhaps the best-known and most widely used learning methods in data mining applications. The reasons for their popularity lie in their conceptual simplicity, ease of usage, computational speed, robustness with respect to missing data and outliers and, most of all, the interpretability of the rules they generate. To separate the observations belonging to different classes, methods based on trees obtain simple and explanatory rules for the relationship existing between the target variable and predictive variables.

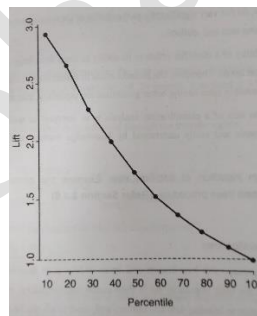
The development of a classification tree corresponds to the training phase of the model and is regulated by a recursive procedure of heuristic nature, based on a divide-and-conquer partitioning scheme referred to as top- down induction

Procedure-Top-down induction of decision trees

- In the initialization phase, each observation is placed in the root node of the tree. The root is included in the list L of active nodes.

-If the list L is empty the procedure is stopped, otherwise a node J belonging to the list L is selected, is removed from the list and is used as the node for analysis.

The optimal rule to split the observations contained in J is then determined, based on an appropriate pre-set criterion. The splitting rule generated in this way is then applied, and descendant nodes are constructed by subdividing the observations contained in J. For each descendant node the conditions for stopping the subdivision are verified. If these are met, node J becomes a leaf, to which the target class is assigned according to the majority of the observations contained in J. Otherwise, the descendant nodes are added to the list L. Finally, step 2 is repeated.



Lift Chart

Components of the top-down induction of decision trees procedure.

Splitting rules. For each node of the tree it is necessary to specify the criteria used to identify the optimal rule for splitting the observations and for creating the descendant nodes. As shown in the next section, there are several alternative criteria, which differ in the number of descendants, the number of attributes and the evaluation metrics.

Stopping criteria. At each node of the tree different stopping criteria are applied to establish whether the development should be continued recursively or the node should be considered as a leaf. In this case too, various criteria have been proposed, which result in quite different topologies of the generated trees, all other elements being equal.

Pruning criteria. Finally, it is appropriate to apply a few pruning criteria, first to avoid excessive growth of the tree during the development phase (pre-pruning), and then to reduce the number of nodes after the tree has been generated (post-pruning).

5. Write a short note on Bayesian methods.? Assume your own training dataset and predict the class label of an unknown sampling using Naïve Bayesian classification.

area	numin	timein	numout	Pothers	Pmob	Pland	numsms	numserv	numcall	diropt	churner
2	1	1	2	1	4	1	3	2	2	0	1
1	1	3	3	2	4	1	4	2	3	0	0
3	2	1	2	2	4	1	3	2	1	0	0
1	2	3	2	3	4	1	1	2	1	0	0
2	3	4	4	4	1	1	3	2	1	0	0
3	3	4	1	4	2	1	4	3	1	0	0
3	3	3	4	4	3	1	4	3	1	1	0
1	1	1	1	1	3	2	1	1	1	0	1
2	2	2	2	1	3	2	2	3	1	1	1
4	2	1	3	2	3	2	1	2	1	1	1
3	1	1	2	2	2	2	2	2	1	0	1
4	3	4	4	2	1	2	2	4	1	1	1
2	1	1	3	2	3	2	2	4	1	1	0
4	2	1	2	3	2	2	2	2	1	0	0
3	3	4	4	3	2	2	2	4	1	1	0
1	1	2	1	4	2	2	2	2	1	0	0
4	1	1	2	4	2	2	4	2	1	0	1
1	1	1	1	1	1	3	1	1	1	0	0
3	1	1	1	1	1	3	1	1	1	0	1
2	3	4	3	1	1	3	1	1	1	0	1
1	3	3	3	1	2	3	4	2	1	0	0
4	2	2	2	2	2	3	1	1	1	0	1
3	3	2	1	4	1	3	1	1	1	0	0

Consider the data given above. The relative frequencies of the sample attribute values given the target class are as follows

area

$P(\text{area} = 1 \mid \text{churner} = 0) = 5/13$ $P(\text{area} = 1 \mid \text{churner} = 1) = 1/10$

$P(\text{area} = 2 \mid \text{churner} = 0) = 2/13$ $P(\text{area} = 2 \mid \text{churner} = 1) = 3/10$

$P(\text{area} = 3 \mid \text{churner} = 0) = 5/13$ $P(\text{area} = 3 \mid \text{churner} = 1) = 2/10$

$P(\text{area} = 4 \mid \text{churner} = 0) = 1/13$ $P(\text{area} = 4 \mid \text{churner} = 1) = 4/10$

numin

$P(\text{numin} = 1 \mid \text{churner} = 0) = 4/13$ $P(\text{numin} = 1 \mid \text{churner} = 1) = 5/10$

$P(\text{numin} = 2 \mid \text{churner} = 0) = 3/13$ $P(\text{numin} = 2 \mid \text{churner} = 1) = 3/10$

$P(\text{numin} = 3 \mid \text{churner} = 0) = 6/13$ $P(\text{numin} = 3 \mid \text{churner} = 1) = 2/10$

timein

$P(\text{timein} = 1 \mid \text{churner} = 0) = 4/13$ $P(\text{timein} = 1 \mid \text{churner} = 1) = 6/10$

$P(\text{timein} = 2 \mid \text{churner} = 0) = 2/13$ $P(\text{timein} = 2 \mid \text{churner} = 1) = 2/10$

$P(\text{timein} = 3 \mid \text{churner} = 0) = 4/13$ $P(\text{timein} = 3 \mid \text{churner} = 1) = 0$

$P(\text{timein} = 4 \mid \text{churner} = 0) = 3/13$ $P(\text{timein} = 4 \mid \text{churner} = 1) = 2/10$

numout

$P(\text{numout} = 1 \mid \text{churner} = 0) = 4/13$ $P(\text{numout} = 1 \mid \text{churner} = 1) = 2/10$

$P(\text{numout} = 2 \mid \text{churner} = 0) = 3/13$ $P(\text{numout} = 2 \mid \text{churner} = 1) = 5/10$

$P(\text{numout} = 3 \mid \text{churner} = 0) = 3/13$ $P(\text{numout} = 3 \mid \text{churner} = 1) = 2/10$

$P(\text{numout} = 4 \mid \text{churner} = 0) = 3/13$ $P(\text{numout} = 4 \mid \text{churner} = 1) = 1/10$

Pothers

$P(\text{Pothers} = 1 \mid \text{churner} = 0) = 2/13$ $P(\text{Pothers} = 1 \mid \text{churner} = 1) = 5/10$

$P(\text{Pothers} = 2 \mid \text{churner} = 0) = 3/13$ $P(\text{Pothers} = 2 \mid \text{churner} = 1) = 4/10$

$P(\text{Pothers} = 3 \mid \text{churner} = 0) = 3/13$ $P(\text{Pothers} = 3 \mid \text{churner} = 1) = 0$

$P(\text{Pothers} = 4 \mid \text{churner} = 0) = 5/13$ $P(\text{Pothers} = 4 \mid \text{churner} = 1) = 1/10$

Pmob

$P(\text{Pmob} = 1 \mid \text{churner} = 0) = 3/13$ $P(\text{Pmob} = 1 \mid \text{churner} = 1) = 3/10$

$P(\text{Pmob} = 2 \mid \text{churner} = 0) = 5/13$ $P(\text{Pmob} = 2 \mid \text{churner} = 1) = 3/10$

$P(\text{Pmob} = 3 \mid \text{churner} = 0) = 2/13$ $P(\text{Pmob} = 3 \mid \text{churner} = 1) = 3/10$

$P(\text{Pmob} = 4 \mid \text{churner} = 0) = 3/13$ $P(\text{Pmob} = 4 \mid \text{churner} = 1) = 1/10$

Pland

$P(\text{Pland} = 1 \mid \text{churner} = 0) = 6/13$ $P(\text{Pland} = 1 \mid \text{churner} = 1) = 1/10$

$P(\text{Pland} = 2 \mid \text{churner} = 0) = 4/13$ $P(\text{Pland} = 2 \mid \text{churner} = 1) = 6/10$
 $P(\text{Pland} = 3 \mid \text{churner} = 0) = 3/13$ $P(\text{Pland} = 3 \mid \text{churner} = 1) = 3/10$

numsms

$P(\text{numsms} = 1 \mid \text{churner} = 0) = 3/13$ $P(\text{numsms} = 1 \mid \text{churner} = 1) = 5/10$
 $P(\text{numsms} = 2 \mid \text{churner} = 0) = 4/13$ $P(\text{numsms} = 2 \mid \text{churner} = 1) = 3/10$
 $P(\text{numsms} = 3 \mid \text{churner} = 0) = 2/13$ $P(\text{numsms} = 3 \mid \text{churner} = 1) = 1/10$
 $P(\text{numsms} = 4 \mid \text{churner} = 0) = 4/13$ $P(\text{numsms} = 4 \mid \text{churner} = 1) = 1/10$

numserv

$P(\text{numserv} = 1 \mid \text{churner} = 0) = 2/13$ $P(\text{numserv} = 1 \mid \text{churner} = 1) = 4/10$
 $P(\text{numserv} = 2 \mid \text{churner} = 0) = 7/13$ $P(\text{numserv} = 2 \mid \text{churner} = 1) = 4/10$
 $P(\text{numserv} = 3 \mid \text{churner} = 0) = 2/13$ $P(\text{numserv} = 3 \mid \text{churner} = 1) = 1/10$
 $P(\text{numserv} = 4 \mid \text{churner} = 0) = 2/13$ $P(\text{numserv} = 4 \mid \text{churner} = 1) = 1/10$

numcall

$P(\text{numcall} = 1 \mid \text{churner} = 0) = 12/13$ $P(\text{numcall} = 1 \mid \text{churner} = 1) = 9/10$
 $P(\text{numcall} = 2 \mid \text{churner} = 0) = 0$ $P(\text{numcall} = 2 \mid \text{churner} = 1) = 1/10$
 $P(\text{numcall} = 3 \mid \text{churner} = 0) = 1/13$ $P(\text{numcall} = 3 \mid \text{churner} = 1) = 0$

diropt

$P(\text{diropt} = 0 \mid \text{churner} = 0) = 10/13$ $P(\text{diropt} = 0 \mid \text{churner} = 1) = 7/10$
 $P(\text{diropt} = 1 \mid \text{churner} = 0) = 3/13$ $P(\text{diropt} = 1 \mid \text{churner} = 1) = 3/10$

Once the conditional probabilities of each attribute given the target class have been estimated, suppose that we wish to predict the target class of a new observation, represented by the vector $x = (1, 1, 1, 2, 1, 4, 2, 1, 2, 1, 0)$. With this aim in mind, we compute the posterior probabilities $P(x|0)$ and $P(x|1)$:

$P(x|0) = 5/13 * 4/13 * 4/13 * 3/13 * 2/13 * 3/13 * 4/13 * 3/13 * 7/13 * 12/13 * 10/13$
 $= 0.81 * 10^{-5}$

$P(x|1) = 1/10 * 5/10 * 6/10 * 5/10 * 5/10 * 1/10 * 6/10 * 5/10 * 4/10 * 9/10 * 7/10$
 $= 5.67 * 10^{-5}$

Since the relative frequencies of the two classes are given by

$P(\text{churner} = 0) = 13/23 = 0.56$, $P(\text{churner} = 1) = 10/23 = 0.44$,

We have

$P(\text{churner} = 0|x) = P(x|0) P(\text{churner} = 0) = 0.81 * 10^{-5} * 0.56 = 0.46 * 10^{-5}$

$P(\text{churner} = 1|x) = P(x|1) P(\text{churner} = 1) = 5.67 * 10^{-5} * 0.44 = 2.495 * 10^{-5}$

The new example x is then labeled with the class value $\{1\}$, since this is associated with the maximum a posteriori probability.

6. Describe Logistic regression.

Logistic regression is a technique for converting binary classification problems into linear regression ones, by means of a proper transformation.

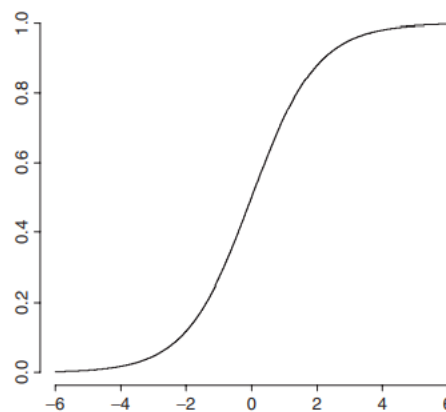
Suppose that the response variable y takes the values $\{0,1\}$, as in a binary classification problem. The logistic regression model postulates that the posterior probability $P(y|x)$ of the response variable conditioned on the vector x follows a logistic function, given by

$$P(y = 0|x) = \frac{1}{1 + e^{w'x}} \dots\dots\dots(1)$$

$$P(y = 1|x) = \frac{e^{w'x}}{1 + e^{w'x}} \dots\dots\dots(2)$$

Here we suppose that the matrix X and the vector w have been extended to include the intercept. The standard logistic function $S(t)$, also known as the sigmoid function, can be found in many applications of statistics in the economic and biological fields and is defined as

$$S(t) = \frac{1}{1 + e^{-t}}$$



Graph of the standard logistic function (sigmoid)

The function $S(t)$ has the graphical shape in the above graph.

By inverting expressions (1) and (2), we observe that the logarithm of the ratio between the conditional probabilities of the two classes depends linearly on the predictive variables, that is

$$\log \frac{P(y = 1|x)}{P(y = 0|x)} = w'x$$

Consequently, by setting

$$z = \log \frac{P(y = 1|x)}{P(y = 0|x)}$$

In general, logistic regression models present the same difficulties described in connection with regression models, from which they derive. To avoid multi-collinearity phenomena that jeopardize the significance of the regression coefficients it is necessary to proceed with attribute selection. Moreover, the accuracy of logistic regression models is in most cases lower than that obtained using other classifiers and usually requires a greater effort for the development of the model. Finally, it appears computationally cumbersome to treat large datasets, both in terms of number of observations and number of attributes.

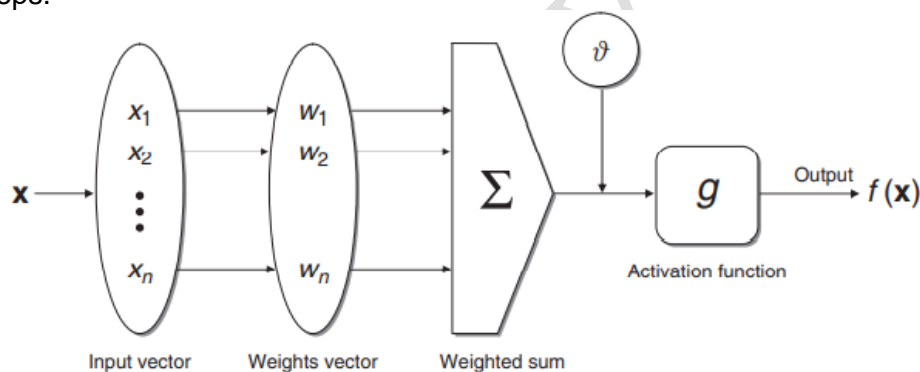
7. Describe neural networks/Explain the ‘Rosenblatt perceptron’ form of neural network with diagram

Neural networks are intended to simulate the behavior of biological systems composed of neurons. A neural network is an oriented graph consisting of nodes, which in the biological analogy represent neurons, connected by arcs, which correspond to dendrites and synapses. Each arc is associated with a weight, while at each node an activation function is defined which is applied to the values received as input by the node along the incoming arcs, adjusted by the weights of the arcs. The training stage is performed by analysing in sequence the observations contained in the training set one after the other and by modifying at each iteration the weights associated with the arcs.

The Rosenblatt perceptron:

The perceptron is the simplest form of neural network and corresponds to a single neuron that receives as input the values (x_1, x_2, \dots, x_n) along the incoming connections, and returns an output value $f(x)$. The input values coincide with the values of the explanatory attributes, while the output value determines the prediction of the response variable y . Each of the n input connections is associated with a weight w_j . An activation function g and a constant ϑ , called the distortion, are also assigned.

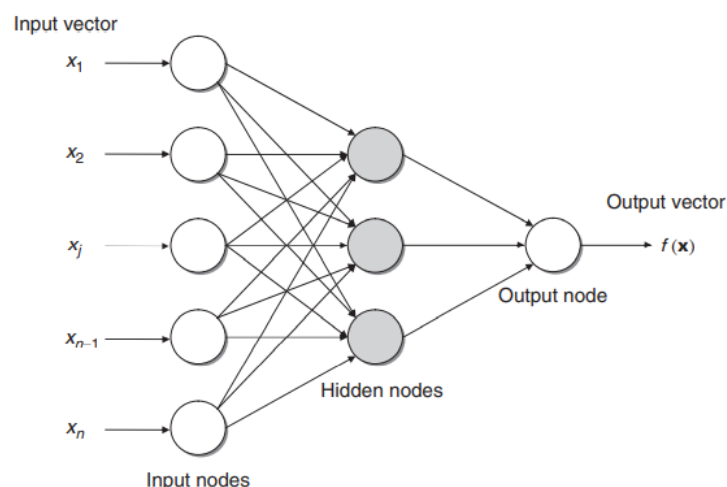
Suppose that the values of the weights and the distortion have already been determined during the training phase. The prediction for a new observation x is then derived by performing the following steps.



Operation of a single unit in a neural network

Multi-level feed-forward networks:

A multi-level feed-forward neural network is a more complex structure than the perceptron, since it includes the following components.



Example of neural network

	<p>Input nodes: The purpose of the input nodes is to receive as input the values of the explanatory attributes for each observation. Usually, the number of input nodes equals the number of explanatory variables.</p> <p>Hidden nodes: Hidden nodes apply given transformations to the input values inside the network. Each node is connected to incoming arcs that go from other hidden nodes or from input nodes, and it is connected with outgoing arcs to output nodes or to other hidden nodes.</p> <p>Output nodes: Output nodes receive connections from hidden nodes or from input nodes and return an output value that corresponds to the prediction of the response variable. In classification problems, there is usually only one output node.</p> <p>Each node of the network basically operates as a perceptron, in the sense that given weights are associated with the input arcs, while each node is associated with a distortion coefficient and an activation function. In general, the activation function may assume forms that are more complex than the sign function $\text{sign}(\cdot)$, such as a linear function, a sigmoid or a hyperbolic tangent.</p> <p>One of the strengths of neural networks is that they are a learning mechanism applicable to both classification and regression problems. Furthermore, they perform attribute selection automatically, since irrelevant or redundant variables can be excluded from the analysis by looking at the coefficients assuming negligible values. However, neural networks require very long times for model training, provide results with modest interpretability that are dependent upon the order in which the examples are analysed, and also present a lower robustness with respect to data affected by noise.</p>
--	--

8.	Write short note on: Clustering methods/What is taxonomy of clustering method? / What are the characteristics of clustering method?
----	--

	<p>Clustering methods:</p> <p>The aim of clustering models is to subdivide the records of a dataset into homogeneous groups of observations, called clusters, so that observations belonging to one group are similar to one another and dissimilar from observations included in other groups. Grouping objects by affinity is a typical reasoning pattern applied by the human brain. Also for this reason clustering models have long been used in various disciplines, such as social sciences, biology, astronomy, statistics, image recognition, processing of digital information, marketing and data mining. Clustering models are useful for various purposes. In some applications, the clusters generated may provide a meaningful interpretation of the phenomenon of interest.</p> <p>Characteristics of clustering methods:</p> <p>Clustering methods must fulfil a few general requirements, as indicated below.</p> <p>Flexibility: Some clustering methods can be applied to numerical attributes only, for which it is possible to use the Euclidean metrics to calculate the distances between observations. However, a flexible clustering algorithm should also be able to analyse datasets containing categorical attributes. Algorithms based on the Euclidean metrics tend to generate spherical clusters and have difficulty in identifying more complex geometrical forms.</p> <p>Robustness: The robustness of an algorithm manifests itself through the stability of the clusters generated with respect to small changes in the values of the attributes of each observation. This property ensures that the given clustering method is basically unaffected by the noise possibly existing in the data. Moreover, the clusters generated must be stable with respect to the order of appearance of the observations in the dataset.</p> <p>Efficiency: In some applications the number of observations is quite large and therefore clustering algorithms must generate clusters efficiently in order to guarantee reasonable computing times for large problems. In the case of massive datasets, one may also resort to the extraction of samples of reduced size in order to generate clusters more efficiently. However, this approach inevitably implies a lower robustness for the clusters so generated. Clustering algorithms must also prove efficient with respect to the number of attributes existing in the dataset.</p>
--	---

	<p>Taxonomy of Clustering Methods: The different types of Clustering based on the logic are partition methods, hierarchical methods, density based methods and grid methods.</p> <p>Types of Clustering:</p> <ol style="list-style-type: none"> 1. Partition methods 2. Hierarchical methods 3. Density based methods 4. Grid methods <p>1. Partition methods Partition methods, is a division of the given dataset into a predetermined number K of non-empty subsets. They generate a spherical or at most convex shape after grouping.</p> <p>2. Hierarchical methods In Hierarchical methods, subset is divided into tree structure. It categorized clusters by different homogeneity thresholds. Predetermined clusters are not required.</p> <p>3. Density-based methods Hierarchical and partition methods are founded on the distance between observations. Density-based methods determine clusters from the number of observations locally falling in a neighbourhood of each observation. For each member which belongs to a specific cluster, a neighbourhood with a specified diameter should contain a number of observations which should not be less than a minimum threshold value. Density-based methods identify clusters of non-convex shape which helps them to isolate any possible outliers.</p> <p>4. Grid methods Grid methods obtain a grid structure consisting of cells. The grid structure is achieved to reduce computing times, despite a lower accuracy in the clusters generated.</p>
9.	<p>What is clustering attributes? Write a short note on:</p> <ol style="list-style-type: none"> A. Binary Attribute B. Nominal Attribute C. Ordinal Attribute
	<p>Clustering Attribute: An attribute is a data field, which represents a characteristic or feature of a data object. The nouns attribute, dimension, feature, and variable are commonly recognized as attribute in literature. In data warehousing attributes are referred as dimension. In Machine learning literature it is referred as feature, while statisticians call this term as variable. Data mining and database professionals commonly use the term attribute. Attributes describing a customer object can include, for example, customer ID, name, and address. Univariate distribution involves only one attribute. The distribution of data having two attributes is known as bivariate. The type of an attribute is determined by the set of possible values the attribute can have. Attributes can be nominal, binary, ordinal, or numeric.</p> <ol style="list-style-type: none"> 1. Binary Attributes: Nominal attribute is treated as binary attribute. It has two categories or states 0 or 1. 0 means attribute is absent and 1 means it is present. Binary attributes are referred to as Boolean as two states correspond to true and false. 1 means that it is present. E.g. Smoker describing a patient object, 1 indicates that the patient smokes, while 0 indicates that the patient does not. A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unlike. The higher the similarity value, the greater the similarity between objects. (Typically, a value of 1 indicates complete similarity, that is, that the objects are identical.) A dissimilarity measure works the opposite way. It returns a value of 0 if the objects are the same (and therefore, far from being dissimilar. The higher the dissimilarity value,

	<p>the more dissimilar the two objects are.</p> <p>2. Nominal Attribute: Nominal attributes means "relating to names." Nominal attribute are symbols or names of things! Each value denotes some kind of category, code, or state. Nominal attributes are also referred as categorical. In computer science, the values are also known as enumerations. Nominal attributes. Suppose that Hair color and Marital status are two attributes describing person objects. In our application, possible values for Hair color are black, brown blond, red, auburn, grey, and white. It is symmetric attribute where the value is greater than 2. We use similarity coefficient in extended form, $\text{dist}(i,j) = (n-f)/n$ Where, f is the number of attributes in which observations i and j take the same value.</p> <p>3. Ordinal Attribute : Values of ordinal attribute has possible values and have a meaningful order or ranking among them. The magnitude between consecutive values is not known. Suppose that Drink size corresponds to the size of drinks available at a restaurant. This ordinal attribute has three possible values - small, medium, and large. However, we cannot tell from the values how much bigger, say, a medium is from a large. Ordinal variable can be discrete or continuous. Order is important and can be treated like interval scaled. Replace ordinal variables value by its rank $r \in \{1, \dots, M_f\}$ Map the range of variable [0,1].</p> $Z_d = \frac{r_r - 1}{M_f - 1}$
10.	<p>Differentiate between following clustering methodologies:</p> <p>A. Partitioning method. B. Hierarchical method.</p>
	<p>Partitioning method: Given a dataset D of m observations, each represented by a vector in n-dimensional space, partition methods construct a subdivision of D into a collection of non-empty subsets $C = \{C_1, C_2, \dots, C_K\}$, where $K \leq m$. In general, the number K of clusters is predetermined and assigned as an input to partition algorithms. The clusters generated by partition methods are usually exhaustive and mutually exclusive, in the sense that each observation belongs to one and only one cluster. There are, however, fuzzy partition methods that assign each observation to different clusters according to a specific proportion. Partition methods start with an initial assignment of the m available observations to the K clusters. Then, they iteratively apply a reallocation technique whose purpose is to place some observations in a different cluster, so that the overall quality of the subdivision is improved. Although alternative measures of the clustering quality can be used, all the various criteria tend to express the degree of homogeneity of the observations belonging to the same cluster and their heterogeneity with respect to the records included in other clusters. Partition algorithms usually stop when during the same iteration no reallocation occurs, and therefore the subdivision appears stable with respect to the evaluation criterion chosen. Partition methods are therefore of a heuristic nature, in the sense that they are based on a myopic logic typical of the class of so-called greedy methods, and at each step they make the choice that locally appears the most advantageous. Operating in this way, there is no guarantee that a globally optimal clustering will be reached, but only that a good subdivision will be obtained, at least for the majority of the datasets. The K-means method and the K-medoids method, which will be described next, are two of the best-known partition algorithms. They are rather efficient clustering methods that are effective in determining clusters of spherical shape.</p>

Hierarchical method:

Hierarchical clustering methods are based on a tree structure. Unlike partition methods, they do not require the number of clusters to be determined in advance. Hence, they receive as input a dataset D containing m observations and a matrix of distances $\text{dist}(x_i, x_k)$ between all pairs of observations. In order to evaluate the distance between two clusters, most hierarchical algorithms resort to one of five alternative measures: minimum distance, maximum distance, mean distance, distance between centroids, and Ward distance. Suppose that we wish to calculate the distance between two clusters C_h and C_f and let z_h and z_f be the corresponding centroids.

Minimum distance:

According to the criterion of minimum distance, also called the single linkage criterion, the dissimilarity between two clusters is given by the minimum distance among all pairs of observations such that one belongs to the first cluster and the other to the second cluster, that is,

$$\text{dist}(C_h, C_f) = \min_{\substack{x_i \in C_h \\ x_k \in C_f}} \text{dist}(x_i, x_k)$$

Maximum distance:

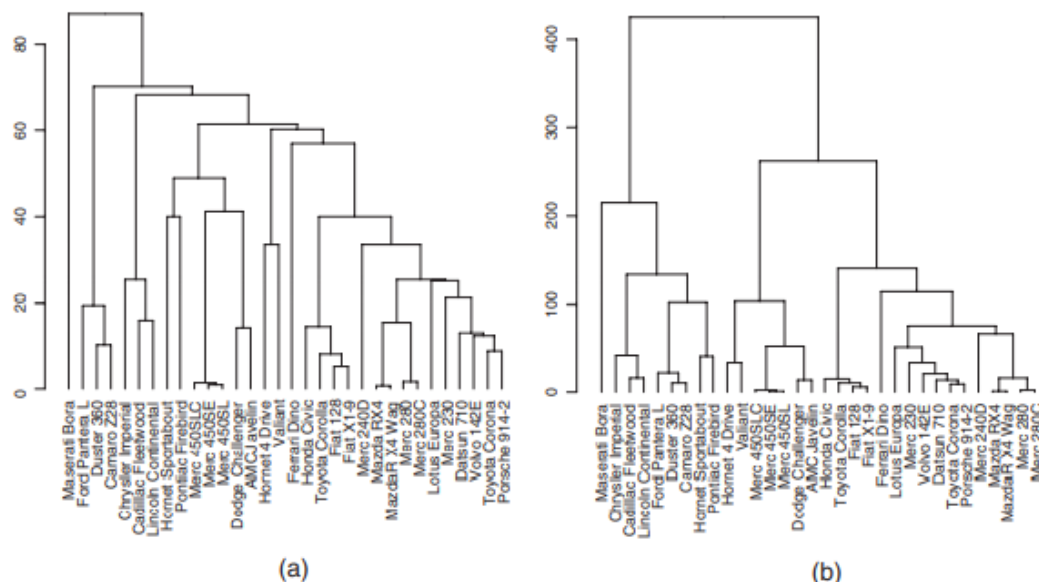
According to the criterion of maximum distance, also called the complete linkage criterion, the dissimilarity between two clusters is given by the maximum distance among all pairs of observations such that one belongs to the first cluster and the other to the second cluster, that is,

$$\text{dist}(C_h, C_f) = \max_{\substack{x_i \in C_h \\ x_k \in C_f}} \text{dist}(x_i, x_k)$$

Mean distance:

The mean distance criterion expresses the dissimilarity between two clusters via the mean of the distances between all pairs of observations belonging to the two clusters, that is,

$$\text{dist}(C_h, C_f) = \frac{\sum_{x_i \in C_h} \sum_{x_k \in C_f} \text{dist}(x_i, x_k)}{\text{card}\{C_h\} \text{card}\{C_f\}}$$



Dendrograms for an agglomerative hierarchical algorithm applied to the mtcars dataset with (a) the mean Euclidean distance and (b) the maximum Euclidean distance

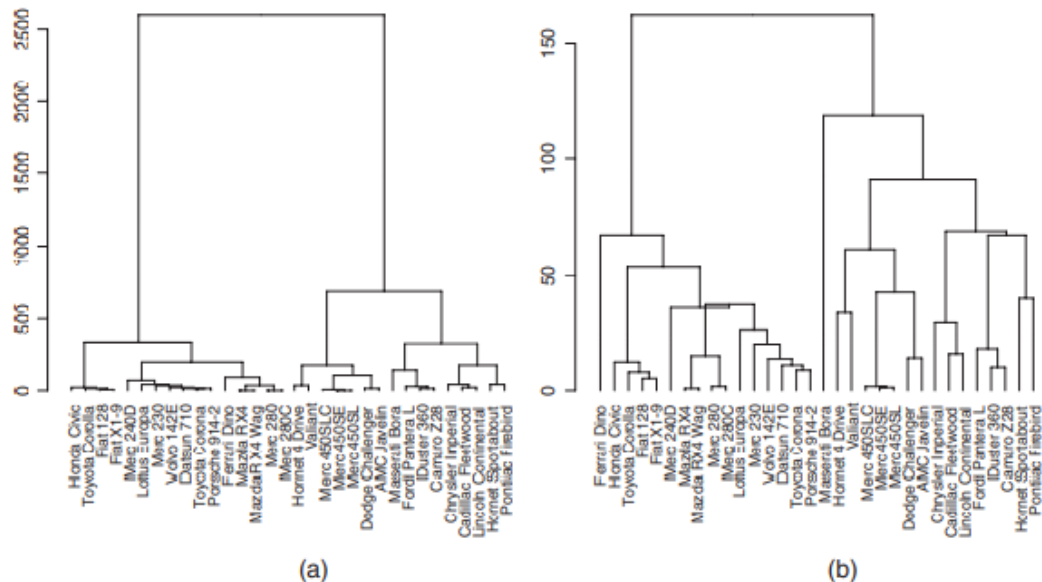
Distance between centroids:

The criterion based on the distance between centroids determines the dissimilarity between two clusters through the distance between the centroids representing the two clusters, that is,

$$\text{dist}(C_h, C_f) = \text{dist}(z_h, z_f)$$

Ward distance:

The criterion of Ward distance, based on the analysis of the variance of the Euclidean distances between the observations, is slightly more complex than the criteria described above. Indeed, it requires the algorithm to first calculate the sum of squared distances between all pairs of observations belonging to a cluster. Methods based on the Ward distance tend to generate a large number of clusters, each containing a few observations.



Dendrograms for an agglomerative hierarchical algorithm applied to the mtcars dataset with (a) the Ward distance and (b) the distance of centroids

Hierarchical methods can be subdivided into two main groups: agglomerative and divisive methods.

1. Agglomerative hierarchical methods:

Agglomerative methods are bottom-up techniques in which each single observation initially represents a distinct cluster. These clusters are then aggregated during subsequent iterations, deriving clusters of increasingly larger cardinalities. The algorithm is stopped when a single cluster including all the observations has been reached. It is then necessary for the user to decide on a cut point and thus determine the number of clusters.

2. Divisive hierarchical methods:

Divisive algorithms are the opposite of agglomerative methods, in that they are based on a top-down technique, which initially places all the observations in a single cluster. This is then subdivided into clusters of smaller size, so that the distances between the generated subgroups are minimized. The procedure is repeated until clusters containing a single observation are obtained, or until an analogous stopping condition is met.

11. Write k-means algorithm for clustering. / Explain k-means method.

K-means algorithm:

- K means clustering is an algorithm used to classify or group the objects based on features or attributes. K is positive integer. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.
- The algorithm assumes two clusters, and each individual's scores include two variables.
- In non-hierarchical clustering such as the k-means algorithm. The relationship between clusters is undetermined. Distance functions such as Manhattan and Euclidian distance functions, are used to determine similarity.
- The K-means algorithm receives as input a dataset D, a number K of clusters to be generated and a function $\text{dist}(x_i, x_k)$ that expresses the inhomogeneity between each pair of observations, or equivalently the matrix D of distances between observations.
- Given a cluster C_h , $h = 1, 2, \dots, K$, the centroid of the cluster is defined as the point z_h having coordinates equal to the mean value of each attribute for the observations belonging to that cluster, that is,

$$z_{hj} = \frac{\sum_{x_i \in C_h} x_{ij}}{\text{card}\{C_h\}}.$$

K-means algorithm

1. During the initialization phase, K observations are arbitrarily chosen in D as the centroids of the clusters.
2. Each observation is iteratively assigned to the cluster whose centroid is the most similar to the observation, in the sense that it minimizes the distance from the record.
3. If no observation is assigned to a different cluster with respect to the previous iteration, the algorithm stops.
4. For each cluster, the new centroid is computed as the mean of the values of the observations belonging to the cluster, and then the algorithm returns to step 2

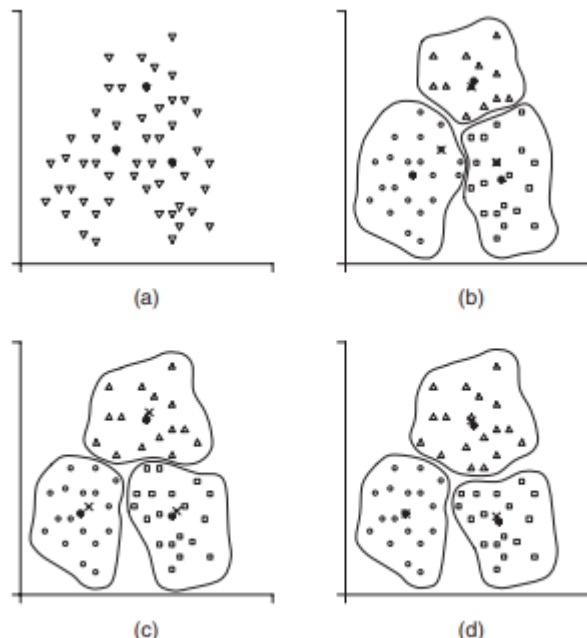
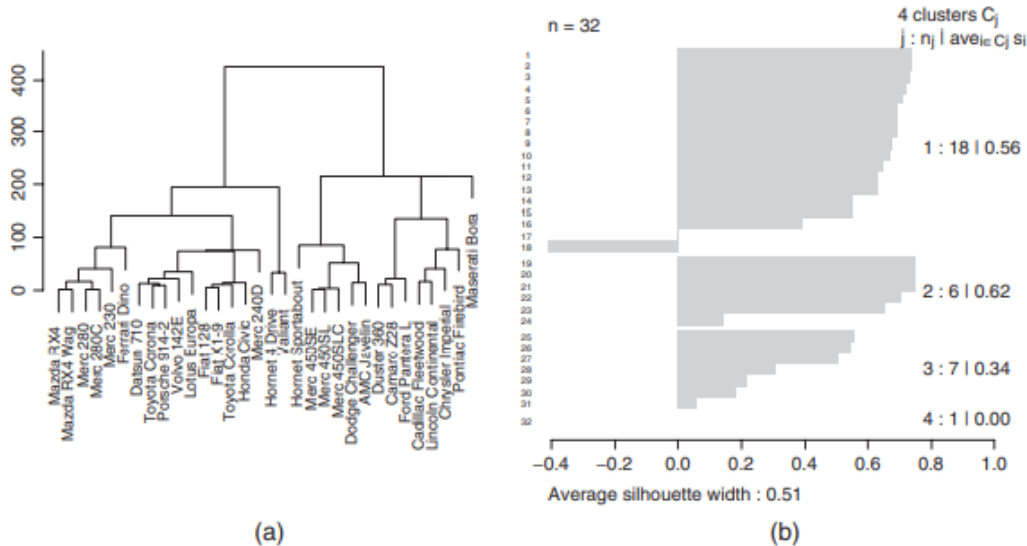


Figure 12.2 An example of application of the K-means algorithm

12. Explain evaluation of clustering model.

For supervised learning methods, such as classification, regression and time series analysis, the evaluation of the predictive accuracy is part of the development process of a model and is based on specific numerical indicators. The same does not apply to clustering methods and, more generally, to unsupervised learning models. Even though the absence of a target attribute makes the evaluation of an unsupervised model less direct and intuitive, it is possible to define reasonable measures of quality and significance for clustering methods.



Dendrograms for a divisive hierarchical algorithm applied to the mtcars dataset (a) with the mean Euclidean distance and (b) the corresponding silhouette with four clusters.

To evaluate a clustering method it is first necessary to verify that the clusters generated correspond to an actual regular pattern in the data. It is therefore appropriate to apply other clustering algorithms and to compare the results obtained by different methods. In this way it is also possible to evaluate if the number of identified clusters is robust with respect to the different techniques applied. At a subsequent phase it is recommended to calculate some performance indicators. Let $C = \{C_1, C_2, \dots, C_K\}$ be the set of K clusters generated. An indicator of homogeneity of the observations within each cluster C_h is given by the cohesion, defined as

$$\text{coh}(C_h) = \sum_{\substack{x_i \in C_h \\ x_k \in C_h}} \text{dist}(x_i, x_k)$$

The overall cohesion of the partition C can therefore be defined as

$$\text{coh}(C) = \sum_{C_h \in C} \text{coh}(C_h).$$

One clustering is preferable over another, in terms of homogeneity within each cluster, if it has a smaller overall cohesion. An indicator of inhomogeneity between a pair of clusters is given by the separation, defined as

$$\text{sep}(C_h, C_f) = \sum_{\substack{x_i \in C_h \\ x_k \in C_f}} \text{dist}(x_i, x_k).$$

Again the overall separation of the partition C can be defined as

$$\text{sep}(C) = \sum_{\substack{C_h \in C \\ C_f \in C}} \text{sep}(C_h, C_f)$$

One clustering is preferable over another, in terms of inhomogeneity among all clusters, if it has a greater overall separation. A further indicator of the clustering quality is given by the silhouette coefficient, which involves a combination of cohesion and separation.

UNIT 4

1. Explain Relational marketing and various factor associated with it.
2. Write a short note on market basket analysis.
3. Explain the concept of acquisition.
4. What is use of web mining methods? What are the different purposes of web mining?
5. Explain sales force management and various factor associated with it.
6. What is supply chain optimization?
7. Explain optimization logistics planning in Logistic and production models.
8. Explain "tactical planning" optimization model for logistics planning.
9. List Revenue management systems. Explain any one in detail.
10. List and explain efficiency measures associated with Data Envelopment analysis./Describe Efficiency measures.
11. Write a short note on efficient frontier.
12. Explain the Charnes–Cooper–Rhodes (CCR) model.
13. Explain basic factors associated with Identification of good operating practices.

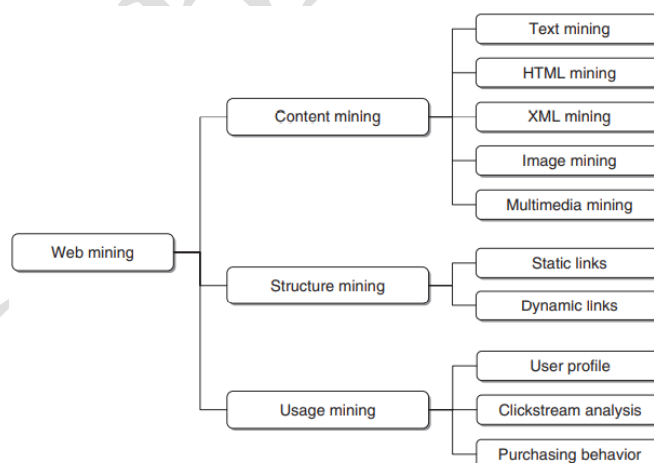
1.	Explain Relational marketing and various factor associated with it.
	<p>We have noticed that whenever a mobile company is about to launch a new device into the market a survey is done by the company so that they get different opinions from their customers, which helps them to enhance the functionality provided by that device. And it is not only about a mobile phone, when you visit a restaurant waiters get the feedback forms along with the bills wherein the customers have to rate the restaurant in different aspects so that they improvise themselves.</p> <p>Almost all the companies study the behaviour and the feedbacks given by the customers and try to inculcate the features that are been required by the customers into their device with a reasonable and effective cost price so that the customers are attracted towards the product and thus sale of the company is increased.</p> <p>Most of the e-commerce company store huge database which have collective information about their customers and the data regarding their previous purchases which helps the company to provide options to its customers which are more likely to be liked by the customers again resulting in growth in the sales of the customers.</p> <p>The strategy that is been followed in relational marketing is to start, strengthen, objectify and maintain the relationship between the customers, stakeholders and the company, which is been presented by the customers, analysis is done, planning is done accordingly, executed and evaluated to achieve the objectives.</p> <p>Relational Marketing evolved and became popular in late 1990s to increase customer's satisfaction so that the competitive advantage is achieved.</p> <p>Initially this approach was initiated by companies providing financial and telecommunication services and later on implemented by almost all the companies wherein they are more concern about what the customer actually needs and accordingly implement the same into their respective products so as to sustain the competitive market.</p>
2.	Write a short note on market basket analysis.
	<p>The purpose of market basket analysis is to gain insight from the purchases made by customers in order to extract useful knowledge to plan marketing actions. It is mostly used to analyze purchases in the retail industry and in e-commerce activities, and is generally amenable to unsupervised learning problems. It may also be applied in other domains to analyze the purchases made using credit cards, the complementary services activated by mobile or fixed telephone customers, the policies or the checking accounts acquired by a same household.</p> <p>The data used for this purpose mostly refer to purchase transactions, and can be associated with the time dimension if the purchaser can be tracked through a loyalty card or the issue of an invoice. Each transaction consists of a list of purchased items. This list is called a basket, just like the baskets available at retail points of sale. If transactions cannot be connected to one another, say because the purchaser is unknown, one may then apply association rules to extract interesting correlations between the purchases of groups of items. The rules extracted in this way can then be used to support different decision-making processes, such as assigning the location of the items on the shelves, determining the layout of a point of sale, identifying which items should be included in promotional flyers, advertisements or coupons distributed to customers.</p> <p>Clustering models are also useful in determining homogeneous groups of items, once an incidence matrix X has been created for the representation of the dataset, where the rows correspond to the transactions and the columns to the items. If customers are individually identified and traced, besides the above techniques it is also possible to develop further analyses that take into account the time dimension of the purchases.</p>

3. Explain the concept of acquisition.

Even if retention is the important aspect of relational marketing strategies acquisition is also an important factor for some of the companies.
It is an process which requires identification of new prospects which are said to be potential customers which can be or may be partially or completely unaware about the products or services that are been provided by the company for did not require this products or services in the past and now are in need of one or the might also be customers of the competitors who are hunting for better services or the other case would be that the customer has switched from your company to the competitor.
Once the company has identified the prospects it is important to assign acquisition campaign with high profitability to both the prospects and the company with various levels marketing strategies along with the marketing resources available with the company.
Traditional marketing strategies are were the advertising and campaign is based on the earlier pools taken from the public in order to enhance the quality of products and services that are been provided which is been fed into data mart to derive classification rules which provides characteristics for the profiles of acquisition.

4. What is use of web mining methods? What are the different purposes of web mining?

The web is a critical channel for the communication and promotion of a company's image. Moreover, e-commerce sites are important sales channels. Hence, it is natural to use web mining methods in order to analyze data on the activities carried out by the visitors to a website. Web mining methods are mostly used for three main purposes: content mining, structure mining and usage mining.



Content mining:

Content mining involves the analysis of the content of web pages to extract useful information. Search engines primarily perform content mining activities to provide the links deemed interesting in relation to keywords supplied by users. Content mining methods can be traced back to data mining problems for the analysis of texts, both in free format or HTML and XML formats, images and multimedia content. Each of these problems is in turn dealt with using the learning models described in previous chapters. For example, text mining analyses are usually handled as multicategory classification problems, where the target variable is the subject category to which the text refers, while explanatory variables correspond to the meaningful words contained in the text. Once it has been converted into a classification problem, text

mining can be approached using the methods. Text mining techniques are also useful for analyzing the emails received by a support center. Notice that the input data for content mining analyses are easily retrievable, at least in principle, since they consist of all the pages that can be visited on the Internet.

Structure mining:

The aim of this type of analysis is to explore and understand the topological structure of the web. Using the links presented in the various pages, it is possible to create graphs where the nodes correspond to the web pages and the oriented arcs are associated with links to other pages. Results and algorithms from graph theory are used to characterize the structure of the web, that is, to identify areas with a higher density of connections, areas disconnected from others and maximal cliques, which are groups of pages with reciprocal links. In this way, it is possible to pinpoint the most popular sites, or to measure the distance between two sites, expressed in terms of the lowest number of arcs along the paths that connect them in the links graph. Besides analyses aimed at exploring the global structure of the web, it is also possible to carry out local investigations to study how a single website is articulated. In some investigations, the local structure of websites is associated with the time spent by the users on each page, to verify if the organization of the site suffers from inconsistencies that jeopardize its effectiveness. For example, a page whose purpose is to direct navigation on the site should be viewed by each user only briefly. Should this not be the case, the page has a problem due to a possible ambiguity in the articulation of the links offered.

Usage mining:

Analyses aimed at usage mining are certainly the most relevant from a relational marketing standpoint, since they explore the paths followed by navigators and their behaviors during a visit to a company website. Methods for the extraction of association rules are useful in obtaining correlations between the different pages visited during a session. In some instances, it is possible to identify a visitor and recognize her during subsequent sessions. This happens if an identification key is required to access a web page, or if a cookie-enabling mechanism is used to keep track of the sequence of visits. Sequential association rules or time series models can be used to analyze the data on the use of a site according to a temporal dynamic. Usage mining analysis is mostly concerned with clickstreams – the sequences of pages visited during a given session. For e-commerce sites, information on the purchase behavior of a visitor is also available.

5. Explain sales force management and various factor associated with it.

Nowadays almost all the companies have sales department into their organizations and rely on the employees of those department for the sales of product or services that are been offered by the company. Every employee is been given a target and depending upon how the targets are been achieved these employees play an important role in the profit that is been gained by the company.

There are various marketing strategies that are been implemented by the sales department for selling off the product or services.

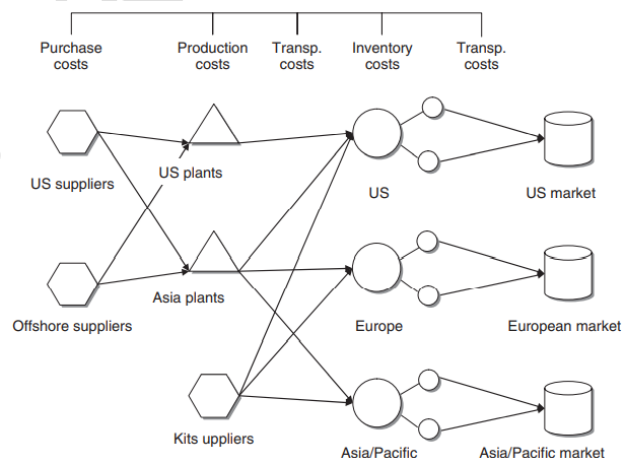
The sales forces is a term coined for all the people and roles along with different tasks and responsibilities that are associated with sales as a process. The basic terms associated with sales forces based on the activities that are been carried out are stated below:

- Residential: This sales activities take place at one or more places which are managed by company supplying products and services from where the customers can purchase, this includes sales at retail shops and wholesale dealers.
- Mobile: In this type of sales the agents of the company go to the customer's house or office to give information about their product or service and also collect the orders. In this category the sale occurs within B2B (Business 2 Business) relationship it can also be encountered in B2C (Business 2 Customer) criteria.

- Telephone: This sales happens on telephonic conversations where the company agents call up the customers and promote the product and also collect the orders. When it comes to mobile sales force there are various problems with it which can be subdivided into few main categories listed below:
 - designing the sales network.
 - planning the agents' activities.
 - contact management.
 - sales opportunity management.
 - customer management.

6. What is supply chain optimization?

In a broad sense, a supply chain may be defined as a network of connected and interdependent organizational units that operate in a coordinated way to manage, control and improve the flow of materials and information originating from the suppliers and reaching the end customers, after going through the procurement, processing and distribution subsystems of a company. The aim of the integrated planning and operations of the supply chain is to combine and evaluate from a systemic perspective the decisions made and the actions undertaken within the various subprocesses that compose the logistic system of a company. Many manufacturing companies, such as those operating in the consumer goods industry, have concentrated their efforts on the integrated operations of the supply chain, even to the point of incorporating parts of the logistic chain that are outside the company, both upstream and downstream. The major purpose of an integrated logistic process is to minimize a function expressing the total cost, which comprises processing costs, transportation costs for procurement and distribution, inventory costs and equipment costs.



An example of global supply chain

The need to optimize the logistic chain, and therefore to have models and computerized tools for medium-term planning and for capacity analysis, is particularly critical in the face of the high complexity of current logistic systems, which operate in a dynamic and truly competitive environment. We are referring here to manufacturing companies that produce a vast array of products and that usually rely on a multicentric logistic system, distributed over several plants and markets, characterized by large investments in highly automated technology, by an

	<p>intensive usage of the available production capacity and by short-order processing cycles. The features of the logistic system we have described reflect the profile of many enterprises operating in the consumer goods industry.</p> <p>In the perspective outlined above, the aim of a medium-term planning process is therefore to devise an optimal logistic production plan, that is, a plan that is able to minimize the total cost, understood as the sum of procurement, processing, storage, distribution costs and the penalty costs associated with the failure to achieve the predefined service level. However, to be implemented in practice, an optimal logistic production plan should also be feasible, that is, it should be able to meet the physical and logical constraints imposed by limits on the available production capacity, specific technological conditions, the structure of the bill of materials, the configuration of the logistic network, minimum production lots, as well as any other condition imposed by the decision makers in charge of the planning process. Optimization models represent a powerful and versatile conceptual paradigm for analyzing and solving problems arising within integrated supply chain planning, and for developing the necessary software. Due to the complex interactions occurring between the different components of a logistic production system, other methods and tools intended to support the planning activity seem today inadequate, such as electronic spreadsheets, simulation systems and planning modules at infinite capacity included in enterprise resource planning software. Conversely, optimization models enable the development of realistic mathematical representations of a logistic production system, able to describe with reasonable accuracy the complex relationships among critical components of the logistic system, such as capacity, resources, plans, inventory, batch sizes, lead times and logistic flows, taking into account the various costs. Moreover, the evolution of information technologies and the latest developments in optimization algorithms mean that decision support systems based on optimization models for logistics planning can be efficiently developed.</p>
7.	Explain optimization logistics planning in Logistic and production models.
	<p>Following are some of the optimization models which are associated with the features of logistic supply chain and logistic production systems.</p> <p>While learning about this models one should understand that real world logistic production systems have more than one element that are been considered so it would be more complex and it will have combination of different features of different elements.</p> <p>Before starting with detailed study of the models some notations that are usually used by these models should be known.</p> <p>In logistic systems I is products denoted by index $i \in I = \{1, 2, \dots, I\}$. Also the planning horizon is been further divided into time intervals T denoted as $t = T = \{1, 2, \dots, T\}$ which is usually of equal length with duration of weeks or months.</p> <p>The manufacturing company have some set of critical resources that are been shared among the companies during the manufacturing process and are also available in limited quantity. These resources may contain manpower, tools, assembly lines, specific fixtures and so on. These critical resources are denoted by R and given as $r \in R = \{1, 2, \dots, R\}$.</p> <p>When even a single critical resource is applicable to the manufacturing process the index value of r is completely omitted to maintain simplicity.</p>

8. Explain "tactical planning" optimization model for logistics planning.

In its simplest form, the aim of tactical planning is to determine the production volumes for each product over the T periods included in the medium-term planning horizon in such a way as to satisfy the given demand and capacity limits for a single resource, and also to minimize the total cost, defined as the sum of manufacturing production costs and inventory costs. We therefore consider the decision variables

P_{it} = units of product i to be manufactured in period t,

I_{it} = units of product i in inventory at the end of period t,

and the parameters

d_{it} = demand for product i in period t,

c_{it} = unit manufacturing cost for product i in period t,

h_{it} = unit inventory cost for product i in period t,

e_i = capacity absorption to manufacture a unit of product i,

b_t = capacity available in period t.

The resulting optimization problem is formulated as follows:

$$\min \sum_{t \in T} \sum_{i \in I} (c_{it} P_{it} + h_{it} I_{it}) \quad \dots\dots\dots(1)$$

$$\text{s.to } P_{it} + I_{i,t-1} - I_{it} = d_{it}, \quad i \in I, t \in T, \quad \dots\dots\dots(2)$$

$$\sum_{i \in I} e_i P_{it} \leq b_t, \quad t \in T, \quad \dots\dots\dots(3)$$

$$P_{it}, I_{it} \geq 0, \quad i \in I, t \in T. \quad \dots\dots\dots(4)$$

Constraints (2) express the balance conditions among production, inventory and demand, by establishing a connection between successive periods along the planning horizon. Inequalities (3) constrain the absorbed capacity not to exceed the available capacity for each period. Model (1) is a linear optimization problem which can be therefore solved efficiently even with a very large number of variables and constraints, of the order of a few million, by means of current state-of-art algorithms and computer technologies.

9. List Revenue management systems. Explain any one in detail.

Revenue management is a policy to manage and its main objective is to maximize the profits for the company by maintaining the balance between demand and supply.

It is usually created for marketing and logistic criteria and has also gained interest in service industries responsible for transport, tourist and hotels.

Eventually it was been accepted by manufacturing and distribution companies. It was expected to grow as the basic idea was related to the revenue and every company thinks about maximizing their profit to the max.

But the revenue management needs to be planned according to the strategies and decision making patterns and models of the company and so it becomes complex when data is feed to it.

10.	List and explain efficiency measures associated with Data Envelopment analysis. / Describe Efficiency measures.
	<p>In data envelopment analysis the units being compared are called decision-making units (DMUs), since they enjoy a certain decisional autonomy. Assuming that we wish to evaluate the efficiency of n units, let $\mathcal{N} = \{1, 2, \dots, n\}$ denote the set of units being compared. If the units produce a single output using a single input only, the efficiency of the jth decision-making unit DMU_j, $j \in \mathcal{N}$, is defined as</p> $\theta_j = \frac{y_j}{x_j},$ <p>in which y_j is the output value produced by DMU_j and x_j the input value used. If the units produce multiple outputs using various input factors, the efficiency of DMU_j is defined as the ratio between a weighted sum of the outputs and a weighted sum of the inputs. Denote by $\mathcal{H} = \{1, 2, \dots, s\}$ the set of production factors and by $\mathcal{K} = \{1, 2, \dots, m\}$ the corresponding set of outputs. If x_{ij}, $i \in \mathcal{H}$, denotes the quantity of input i used by DMU_j and y_{rj}, $r \in \mathcal{K}$, the quantity of output r obtained, the efficiency of DMU_j is defined as</p> $\theta_j = \frac{u_1 y_{1j} + u_2 y_{2j} + \dots + u_m y_{mj}}{v_1 x_{1j} + v_2 x_{2j} + \dots + v_s x_{sj}} = \frac{\sum_{r \in \mathcal{K}} u_r y_{rj}}{\sum_{i \in \mathcal{H}} v_i x_{ij}},$ <p>for weights u_1, u_2, \dots, u_m associated with the outputs and v_1, v_2, \dots, v_s assigned to the inputs. In this second case, the efficiency of DMU_j depends strongly on the system of weights introduced. At different weights, the efficiency value may undergo relevant variations and it becomes difficult to fix a single structure of weights that might be shared and accepted by all the evaluated units. In order to avoid possible objections raised by the units to a preset system of weights, which may privilege certain DMUs rather than others, data envelopment analysis evaluates the efficiency of each unit through the weights system that is best for the DMU itself – that is, the system that allows its efficiency value to be maximized. Subsequently, by means of additional analyses, the purpose of data envelopment analysis is to identify the units that are efficient in absolute terms and those whose efficiency value depends largely on the system of weights adopted.</p>
11.	Write a short note on efficient frontier.
	<p>Efficient frontier:</p> <p>The efficient frontier, also known as production function, expresses the relationship between the inputs utilized and the outputs produced. It indicates the maximum quantity of outputs that can be obtained from a given combination of inputs. At the same time, it also expresses the minimum quantity of inputs that must be used to achieve a given output level. Hence, the efficient frontier corresponds to technically efficient operating methods. The efficient frontier may be empirically obtained based on a set of observations that express the output level obtained by applying a specific combination of input production factors. In the context of data envelopment analysis, the observations correspond to the units being evaluated. Most statistical methods of parametric nature, which are based for instance on the calculation of a regression curve, formulate some prior hypotheses on the shape of the production function. Data envelopment analysis, on the other hand, forgoes any assumptions on the functional form of the efficient frontier, and is therefore nonparametric in character. It only requires that the units being compared are not placed above the production function, depending on their efficiency value.</p>

12. Explain the Charnes–Cooper–Rhodes (CCR) model.

Using data envelopment analysis, the choice of the optimal system of weights for a generic DMU_j involves solving a mathematical optimization model whose decision variables are represented by the weights u_r , $r \in K$, and v_i , $i \in H$, associated with each output and input. Various formulations have been proposed, the best-known of which is probably the Charnes–Cooper–Rhodes (CCR) model. The CCR model formulated for DMU_j takes the form

$$\max \quad \vartheta = \frac{\sum_{r \in K} u_r y_{rj}}{\sum_{i \in H} v_i x_{ij}}, \quad \dots\dots\dots(1)$$

$$\text{s.to} \quad \frac{\sum_{r \in K} u_r y_{rj}}{\sum_{i \in H} v_i x_{ij}} \leq 1, \quad j \in N, \quad \dots\dots\dots(2)$$

$$u_r, v_i \geq 0, \quad r \in K, i \in H. \quad \dots\dots\dots(3)$$

The objective function involves the maximization of the efficiency measure for DMU_j. Constraints (2) require that the efficiency values of all the units, calculated by means of the weights system for the unit being examined, be lower than one. Finally, conditions (3) guarantee that the weights associated with the inputs and the outputs are non-negative. In place of these conditions, sometimes the constraints $u_r, v_i \geq \delta$, $r \in K$, $i \in H$ may be applied, where $\delta > 0$, preventing the unit from assigning a null weight to an input or output. Model (1) can be linearized by requiring the weighted sum of the inputs to take a constant value, for example 1. This condition leads to an alternative optimization problem, the input-oriented CCR model, where the objective function consists of the maximization of the weighted sum of the outputs

$$\max \quad \vartheta = \sum_{r \in K} u_r y_{rj}, \quad \dots\dots\dots(4)$$

$$\text{s.to} \quad \sum_{i \in H} v_i x_{ij} = 1, \quad \dots\dots\dots(5)$$

$$\sum_{r \in K} u_r y_{rj} - \sum_{i \in H} v_i x_{ij} \leq 0, \quad j \in N, \quad \dots\dots\dots(6)$$

$$u_r, v_i \geq 0, \quad r \in K, i \in H. \quad \dots\dots\dots(7)$$

Let ϑ^* be the optimum value of the objective function corresponding to the optimal solution (v^* , u^*) of problem (4). DMU_j is said to be efficient if $\vartheta^* = 1$ and if there exists at least one optimal solution (v^* , u^*) such that $v^* > 0$ and $u^* > 0$. By solving a similar optimization model for each of the n units being compared, one obtains n systems of weights. The flexibility enjoyed by the units in choosing the weights represents an undisputed advantage, in that if a unit turns out to be inefficient based on the most favorable system of weights, its inefficiency cannot be traced back to an inappropriate evaluation process. However, given a unit that scores $\vartheta^* = 1$, it is important to determine whether its efficiency value should be attributed to an actual high-level performance or simply to an optimal selection of the weights structure.

13. Explain basic factors associated with Identification of good operating practices.

Having good operating practices is important as it helps to improve the performance given by unit when compared.

The units that are said to be capable in terms of data envelopment analysis demonstrate to compare and also examples that are associated with other units.

Also between all the most efficient units there might be some which will help to improve the existing ability. It is important to search for most capable unit so that the ability of existing operating practices is improved.

So to identify great operating practices the units that are actually capable need to be recognized and their score also demonstration depends on system of weights that are selected.

To distinguish between these units we can use different methods like cross-efficiency analysis, evaluation of virtual inputs and virtual outputs, and weight restrictions.

Cross-Efficiency Analysis

Cross efficiency analysis is done with the help of efficiency matrix that gives informatics about the nature of weights systems which are been implemented by units for their ability calculation.

The square efficiency matrix contains multiple rows and columns that have units that are been compared. The element θ_{ij} of matrix denotes ability of DMU, calculated with optimal weights structure for DMU_j and θ_{ij} ability of DMU, which is evaluated using optimal weights.

If DMU is efficient i.e. $\theta = 1$ even if it shows behaviour which is to be sustained to special dimension along with units the ability value in column related to DMU_j that should be less than 1.

The quantities of interest can be derived from efficiency matrix. In which first is the average ability which is obtained from j^{th} column whereas second is average efficiency obtained by measuring unit of optimal system of weights to other units.

Later is gained by averaging values in rows which is been associated with units that are been examined.

The difference between θ_{ij} and DMU_j and ability gained as average value of j^{th} column gives the result of how much the unit relies on system weights that is been used by units to calculate the process.

If the difference obtained between the two terms is relevant, DMU_j will choose structure that is not been shared by other DMU in order to given all the privilege of analysis for efficient functioning.

Virtual Inputs and Virtual Outputs

Virtual inputs and virtual outputs gives information about importance of every units features for every input and output for the reason to maximize its ability score. And hence allows some specific capability of every unit identified, highlighted and also its weaknesses are been presented at same time. The virtual inputs that are of DMU are said to be the product of inputs that are been used by unit and its interrelated weights.

UNIT 5

1. Define knowledge management. What are data, information and knowledge?/ Write a short note on approaches Knowledge Management.
2. What is meant by knowledge management system? / Describe the knowledge management system (KMS) cycle.
3. Explain the role of people in knowledge management.
4. Explain organizational learning and Transformation.
5. Explain knowledge management activities in brief
6. What is information technology in Knowledge Management?
7. Write steps involved in knowledge management system implementation.
8. What is Artificial intelligence? / List and explain characteristics of artificial intelligence.
9. Describe how AI and intelligent agents support knowledge management. Relate XML to knowledge management and knowledge portals. / Compare and contrast between Artificial intelligence versus Natural intelligence.
10. What are expert system? / Explain basic concepts and structure of Expert Systems./ Enlist and explain steps of development of expert system.
11. Explain forward chaining and backward chaining.
12. What are the areas for expert system applications?
13. What is knowledge engineering? Explain the process of knowledge engineering. / Write short note on knowledge engineering.

1.	Define knowledge management. What are data, information and knowledge?
	<p>Knowledge management: Knowledge management is an activity practised by enterprises all over the world. In the process of knowledge management, these enterprises comprehensively gather information using many methods and tools. Then, gathered information is organized, stored, shared, and analysed using defined techniques. The analysis of such information will be based on resources, documents, people and their skills. Properly analysed information will then be stored as 'knowledge' of the enterprise. This knowledge is later used for activities such as organizational decision making and training new staff members. There have been many approaches to knowledge management from early days. Most of early approaches have been manual storing and analysis of information. With the introduction of computers, most organizational knowledge and management processes have been automated. Therefore, information storing, retrieval and sharing have become convenient. Nowadays, most enterprises have their own knowledge management framework in place. The framework defines the knowledge gathering points, gathering techniques, tools used, data storing tools and techniques and analysing mechanism. Knowledge management (KM) is the systematic and active management of ideas, information, and knowledge residing in an organization's employees. Its purposes include effective and efficient problem solving, dynamic learning, strategic planning, and decision making. KM initiatives focus on identifying knowledge, explicating it in such a way that it can be shared in a formal manner, and leveraging its value through reuse Data are facts, measurements, and statistics. Information is organized or processed data that is timely (i.e., inferences from the data are drawn within the time frame of applicability) and accurate (i.e., with regard to the original data). Knowledge is information that is contextual, relevant, and actionable and describe its purpose. Data: Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. For example, for a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions. Information: Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain. For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over ₹100 per week, or the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data. Knowledge: Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems. For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business.</p>

2.	What is meant by knowledge management system? / Describe the knowledge management system (KMS) cycle.
	<p>Knowledge Management System (KMS): Knowledge is a key to success. Knowledge management is one of the most important activities that an organization has to adapt.</p> <ol style="list-style-type: none"> 1. <u>Share and Learn</u>: The sharing of knowledge in order to facilitate learning is the first step in knowledge management life-cycle. Sharing of knowledge is one in which people exchange their views and ideas on a particular domain. 2. <u>Create</u>: Knowledge is created by sharing of ideas by people working in an organization. Better sharing leads to better ideas thereby creating a valuable knowledge repository. 3. <u>Capture and Acquire</u>: Capture and acquisition of knowledge is one in which the knowledge created is collected in huge numbers and stored in a repository. 4. <u>Organize</u>: Organizing is the next step to capturing of knowledge. The captured content is organized using a framework or knowledge model. The model reflects the elements of knowledge and flows that are embedded inherently in the specific processes and culture of organization. 5. <u>Access, Search and Disseminate</u>: The organized knowledge is put in such a way that it could be accessed, searched and disseminated by the users working in the organization. 6. <u>Use and Discover</u>: The last step is to make use of the knowledge acquired in solving problems in real time. <p>As seen above, the key to knowledge management lies in sharing of knowledge. Sharing the knowledge increases the innovation and improves the overall quality of work. Thus, proper knowledge management helps organizations in developing the skill set of employees and improving their overall efficiency at work.</p>
3.	Explain the role of people in knowledge management.
	<p>People are ultimately the holders of knowledge. The goal is to encourage them to not only search for it and improve it for applying it to improving internal processes, but to make them see the benefits of sharing it with the organization, in this context it is important:</p> <ol style="list-style-type: none"> 1. To give people autonomy in their jobs and find new ways to fulfill them. 2. To provide proper storage and sharing of knowledge systems. 3. To empower them and continually train them 4. To keep them motivated 5. To give them adequate remuneration, to ensure their commitment. <p>The manager should always be aware of the fact that decisions made by people can affect the entire organization.</p> <p>That's why your motivation is crucial, that's what will make employees share and replicate the knowledge they accumulate in their activities in the company with colleagues.</p> <p>The worst that can happen is to lose that talent to the competition, along with everything they have learned.</p>

4. Explain organizational learning and Transformation.

Learning organizational:

The learning organisation is an organisation characterised by a deep commitment to learning and education with the intention of continuous improvement. This concept reviews several theories relating to the learning organisation, including some criticism.

Also, it examines some evidence on how learning organisations operate. Learning organisations facilitate collective learning in order to continually improve the capacity to respond to changing demands in the environment. This permeates all organisational activities, structures, processes, climate and values, leading to an enhanced ability to react quickly to opportunities and threats.

Organizational Transformation

Organizational transformation takes place when there is a change in the way the business is done or in the event of a re-engineering or restructuring activity.

Along with the structural changes, the attitude of the employees, their perspectives as well as the culture of the organization undergoes a significant change. It's about re-modelling an organization in its entirety.

There are three key stages for managing organisational transformation along with the critical success factors for managing change at each stage.

Stage 1: Break with the past

- Bring in outsiders. The Board should introduce entrepreneurial outsiders with targeted expertise onto the top management team.

Break with your administrative heritage. Important mechanisms here can be the removal of blockers, rotation of managers, promotion of young managers untainted by the organisational heritage, the utilisation of project teams, the achievement of early successes and designing a suitable bonus/incentive system.

Use aspects of the administrative heritage that help the change process. Not everything that worked in the past needs to be thrown away.

This will vary from company to company. Some may be able leverage a traditional command-and-control management style to achieve more rapid implementation of change; however, in environments where a more democratic leadership style is the norm, it may be more appropriate to leverage other factors, for example, customer relationships, a strong R&D department, or the latent enthusiasm of organisational members for participating in new initiatives. Crisis is also an important lever for organisational change.

Stage 2: Manage the present

Vary your leadership style as appropriate. The top-down approach of Stage I may be still required to break with the past in some parts of the organisation, while other parts may by this stage already have the ability to learn and therefore may be given authority and empowerment to act.

Exploit best practice from your own or other organisations. This will require knowledge acquisition, knowledge internalisation and knowledge dissemination. Reconfigure, divest and integrate resources. This involves everything from streamlining business systems to removing non-aligned employees to consolidating new acquisitions operationally and culturally.

Stage 3: Invest in the future

Empower the organisation. The top management team should delegate to employees as well as motivating and enabling them to act.

Enable the organisation to engage in exploration of new ideas and business practices. You can achieve this by encouraging innovation, trial and experimentation and by developing a culture which encourages informed risk-taking and facilitates learning from mistakes.

Exploration enables the organisation to develop new capabilities fitted to its specific context,

	<p>rather than just importing systems and routines from other contexts. Create new paths. This means creating a deliberate change in direction using new capabilities, whether that be in terms of new products, services, processes or business models.</p> <p>The combination of exploration and path creation will lead you to the "disruptive innovation" that will help you secure sustainable competitive advantage.</p> <p>By going through these stages, organizations can establish new developmental pathways, enhance their strategic flexibility, and react successfully to changes in the environment.</p>
5.	Explain knowledge management activities in brief.
	<p>A winning knowledge management program increases staff productivity, product and service quality, and deliverable consistency by capitalizing on intellectual and knowledge-based assets.</p> <p>Many organizations leap into a knowledge management solution (e.g. document management, data mining, blogging, and community forums) without first considering the purpose or objectives they wish to fulfill or how the organization will adopt and follow best practices for managing its knowledge assets long term. A successful knowledge management program will consider more than just technology, An organization should also consider:</p> <p>People: They represent how you increase the ability of individuals within the organization to influence others with their knowledge.</p> <p>Processes: They involve how you establish best practices and governance for the efficient and accurate identification, management, and dissemination of knowledge.</p> <p>Technology: It addresses how you choose, configure, and utilize tools and automation to enable knowledge management.</p> <p>Structure: It directs how you transform organizational structures to facilitate and encourage cross-discipline awareness and expertise.</p> <p>Culture: It embodies how you establish and cultivate a knowledge-sharing, knowledge- driven culture.</p> <p>The Power of Knowledge Management</p> <p>Implementing a complete knowledge management takes time and money, however, the results can be impressive and risks can be minimized by taking a phased approach that gives beneficial returns at each step.</p> <p>Organizations that have made this kind of investment in knowledge management realize tangible results quickly.</p> <p>They add to their top and bottom lines through faster cycle times, enhanced efficiency, better decision making and greater use of tested solutions across the enterprise.</p>
6.	What is information technology in Knowledge Management?
	<p>KM was initially driven primarily by IT, information technology, and the desire to put that new technology, the Internet, to work and see what it was capable of.</p> <p>That first stage has been described using a horse breeding metaphor as "by the internet out of intellectual capital," the sire and the dam.</p> <p>The concept of intellectual capital, the notion that not just physical resources, capital, and manpower, but also intellectual capital (knowledge) fueled growth and development provided the justification, the framework, and the seed. The availability of the internet provided the tool.</p> <p>As described above, the management consulting community jumped at the new capabilities provided by the Internet, using it first for themselves, realizing that if they shared knowledge</p>

	<p>across their organization more effectively they could avoid reinventing the wheel, underbid their competitors, and make more profit.</p> <p>The central point is that the first stage of KM was about how to deploy that new technology to accomplish more effective use of information and knowledge.</p> <p>The first stage might be described as the "If only Texas Instruments knew what Texas Instruments knew" stage, to revisit a much quoted KM mantra. The hallmark phrase of Stage/ was first "best practices." later replaced by the more politic "lessons learned."</p>
7.	<p>Write steps involved in knowledge management system implementation</p> <p>Steps to Implementation: Implementing a knowledge management program is no easy feat. You will encounter many challenges along the way including many of the following:</p> <ul style="list-style-type: none"> ✓ Inability to recognize or articulate knowledge; turning tacit knowledge into explicit knowledge. ✓ Geographical distance and/or language barriers in an international company. Limitations of information and communication technologies. ✓ Loosely defined areas of expertise. ✓ Internal conflicts (e.g. professional territoriality). Lack of incentives or performance management goals. ✓ Poor training or mentoring programs. ✓ Cultural barriers (e.g. "this is how we've always done it" mentality). <p>The following eight-step approach will enable you to identify these challenges so you can plan for them, thus minimizing the risks and maximizing the rewards. This approach was developed based on logical, tried-and-true activities for implementing any new organizational program. The early steps involve strategy, planning, and requirements gathering while the later steps focus on execution and continual improvement.</p> <p>Step 1: Establish Knowledge Management Program Objectives Before selecting a tool, defining a process, and developing workflows, you should envision and articulate the end state. In order to establish the appropriate program objectives, identify and document the business problems that need resolution and the business drivers that will provide momentum and justification for the endeavor. Provide both short-term and long-term objectives that address the business problems and support the business drivers. Short-term objectives should seek to provide validation that the program is on the right path while long-term objectives will help to create and communicate the big picture.</p> <p>Step 2: Prepare for Change Knowledge management is more than just an application of technology. It involves cultural changes in the way employees perceive and share knowledge they develop or possess. One common cultural hurdle to increasing the sharing of knowledge is that companies primarily reward individual performance. This practice promotes a "knowledge is power" behavior that contradicts the desired knowledge-sharing, knowledge-driven culture end state you are after. Successfully implementing a new knowledge management program may require changes within the organization's norms and shared values; changes that some people might resist or even attempt to quash. To minimize the negative impact of such changes, it's wise to follow an established approach for managing cultural change.</p>

Step 3: Define High-Level Process

To facilitate the effective management of your organization's knowledge assets, you should begin by laying out a high-level knowledge management process.

The process can be progressively developed with detailed procedures and work instructions throughout steps four, five, and six. However, it should be finalized and approved prior to step seven (implementation). Organizations that overlook or loosely define the knowledge management process will not realize the full potential of their knowledge management objectives.

How knowledge is identified, captured, categorized, and disseminated will be ad hoc at best. There are a number of knowledge management best practices, all of which comprise similar activities.

-In general, these activities include knowledge strategy, creation, identification, classification, capture, validation, transfer, maintenance, archival, measurement, and reporting.

Step 4: Determine and Prioritize Technology Needs

Depending on the program objectives established in step one and the process controls and criteria defined in step three, you can begin to determine and prioritize your knowledge management technology needs.

With such a variety of knowledge management solutions, it is imperative to understand the cost and benefit of each type of technology and the primary technology providers in the marketplace.

Don't be too quick to purchase a new technology without first determining if your existing technologies can meet your needs.

You can also wait to make costly technology decisions after the knowledge management program is well underway if there is broad support and a need for enhanced computing and automation.

Step 5: Assess Current State

Now that you've established your program objectives to solve your business problem, prepared for change to address cultural issues, defined a high-level process to enable the effective management of your knowledge assets, and determined and prioritized your technology needs that will enhance and automate knowledge management related activities, you are in a position to assess the current state of knowledge management within your organization.

The knowledge management assessment should cover all five core knowledge management components: people, processes, technology, structure, and culture.

A typical assessment should provide an overview of the assessment, the gaps between current and desired states, and the recommendations for attenuating identified gaps. The recommendations will become the foundation for the roadmap in step six.

Step 6: Build a Knowledge Management Implementation Roadmap

With the current-state assessment in hand, it is time to build the implementation roadmap for your knowledge management program.

But before going too far, you should re-confirm senior leadership's support and commitment, as well as the funding to implement and maintain the knowledge management program.

Without these prerequisites, your efforts will be futile. Having solid evidence of your organization's shortcomings, via the assessment, should drive the urgency rate up. Having a strategy on how to overcome the shortcomings will be critical in gaining leadership's support and getting the funding you will need.

This strategy can be presented as a roadmap of related projects, each addressing specific gaps identified by the assessment.

The roadmap can span months and years and illustrate key milestones and dependencies.

A good roadmap will yield some short-term wins in the first step of projects, which will bolster support for subsequent steps.

	<p>As time progresses, continue to review and evolve the roadmap based upon the changing economic conditions and business drivers.</p> <p>7.Implementation: Implementing a knowledge management program and maturing the overall effectiveness of your organization will require significant personnel resources and funding. Be prepared for the long haul, but at the same time, ensure that incremental advances are made and publicized. As long as there are recognized value and benefits, especially in light of ongoing successes, there should be little resistance to continued knowledge management investments. With that said, it's time for the rubber to meet the road. You know what the objectives are,</p> <p>Step 8: Measure and Improve the Knowledge Management Program How will you know your knowledge management investments are working? You will need a way of measuring your actual effectiveness and comparing that to anticipated results. If possible, establish some baseline measurements in order to capture the before shot of the organization's performance prior to implementing the knowledge management program. Then, after implementation, trend and compare the new results to the old how performance has improved. Don't be disillusioned if the delta is not as large as you would have anticipated. It will take time for the organization to become proficient with the new processes and improvements. Over time, the results should follow suit. When deciding upon the appropriate metrics to measure your organization's progress.</p>
8.	<p>What is Artificial intelligence? / List and explain characteristics of artificial intelligence.</p>
	<p>Artificial Intelligence: Since the invention of computers or machines, their capability to perform various tasks went on growing exponentially. Humans have developed the power of computer systems in terms of their diverse working domains, their increasing speed, and reducing size with respect to time. A branch of Computer Science named Artificial Intelligence pursues creating the computers or machines as intelligent as human beings. According to the father of Artificial Intelligence, John McCarthy, it is "The science and engineering of making intelligent machines, especially intelligent computer programs". Artificial Intelligence is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think. AI is accomplished by studying how human brain thinks, and how humans learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems.</p> <p>Characteristics of Artificial Intelligence: Gaming: AI plays crucial role in strategic games such as chess, poker, tic-tac-toe, etc., where machine un think of large number of possible positions based on heuristic knowledge. Natural Language Processing: It is possible to interact with the computer that understands natural language spoken by humans. Expert Systems: There are some applications which integrate machine, software, and special information to impart reasoning and advising. They provide explanation and advice to the users. Vision Systems : These systems understand, interpret, and comprehend visual input on the computer. For example, ✓ A spying aeroplane takes photographs, which are used to figure out spatial information or map of the areas.</p>

	<ul style="list-style-type: none"> ✓ Doctors use clinical expert system to diagnose the patient. ✓ Police use computer software that can recognize the face of criminal with the stored portrait made by forensic artist. <p>Speech Recognition: Some intelligent systems are capable of hearing and comprehending the language in terms of sentences and their meanings while a human talk to it. It can handle different accents, slang words noise in the background, change in human's noise due to cold, etc.</p> <p>Handwriting Recognition: The handwriting recognition software reads the text written on paper by a pen or on screen by a stylus. It can recognize the shapes of the letters and convert it into editable text.</p> <p>Intelligent Robots: Robots are able to perform the tasks given by a human. They have sensors to detect physical data from the real world such as light, heat, temperature, movement, sound, bump, and pressure. They have efficient processors, multiple sensors and huge memory, to exhibit intelligence. In addition, they are capable of learning from their mistakes and they can adapt to the new environment.</p>
--	--

9.	<p>Describe how AI and intelligent agents support knowledge management. Relate XML to knowledge management and knowledge portals. / Compare and contrast between Artificial intelligence versus Natural intelligence</p>
----	---

	<p>Intelligence can be defined as a general mental ability for reasoning, problem-solving, and learning. Because of its general nature, intelligence integrates cognitive functions such as perception, attention, memory, language, or planning.</p> <p>On the basis of this definition, intelligence can be reliably measured by standardized tests with obtained scores predicting several broad social outcomes such as educational achievement, job performance, health, and longevity. So let's study the differences between Artificial Intelligence and Human Intelligence in a detail.</p> <p>Artificial intelligence Artificial Intelligence is the study and design of Intelligent agent. These intelligent agents have the ability to analyse the environments and produce actions which maximize success.</p> <p>AI research uses tools and insights from many fields, including computer science, psychology, philosophy, neuroscience, cognitive science, linguistics, operations research economics, control theory, probability, optimization and logic.</p> <p>AI research also overlaps with tasks such as robotics, control systems, scheduling, data mining, logistics, speech recognition, facial recognition and many others.</p> <p>Human Intelligence Human Intelligence is defined as the quality of the mind that is made up of capabilities to learn from past experience, adaptation to new situations, handling of abstract ideas and the ability to change his/her own environment using the gained knowledge.</p> <p>Human Intelligence can provide several kinds of information. It can provide observations during travel or other events from travellers, refugees, escaped friendly POWs, etc.</p> <p>It can provide data on things about which the subject has specific knowledge, which can be another human subject, or, in the case of defectors and spies, sensitive information to which they had access. Finally, it can provide information on interpersonal relationships and networks of interest.</p>
--	--

	<p>Key Differences between Artificial Intelligence and Human Intelligence</p> <ol style="list-style-type: none"> 1. Nature of Existence 2. Memory usage 3. Mode of creation 4. Learning process 5. Dominance <p>1. Nature of Existence Human intelligence revolves around adapting to the environment using a combination of several cognitive processes. The field of Artificial intelligence focuses on designing machines that can mimic human behaviour.</p> <p>2. Memory usage Humans use content memory and thinking whereas, robots are using the built-in instructions, designed by scientists.</p> <p>3. Mode of creation Human intelligence is bigger because its creation of God and artificial intelligence as the name suggests is artificial, little and temporary created by humans. Also, Humans intelligence is the real creator of the artificial intelligence even but they cannot create a human being with superiority.</p> <p>4. Learning process Human intelligence is based on the variants they encounter in life and responses they get which may result in millions of functions overall in their lives. However, for Artificial intelligence is defined or developed for specific tasks only and its applicability on other tasks not be easily possible.</p> <p>5. Dominance Artificial intelligence can beat human intelligence in some specific areas such as in Chess a supercomputer has beaten the human player due to being able to store all the moves played by humans so far and being able to think ahead 10 moves as compared to human players who can think 10 steps ahead but cannot store and retrieve that number of moves in Chess.</p>
10.	<p>What are expert system? / Explain basic concepts and structure of Expert Systems./ Enlist and explain steps of development of expert system.</p>
	<p>Expert systems: Expert Systems (ES) are one of the prominent research domains of AI. It is introduced by the researchers at Stanford University, Computer Science Department. The expert systems are the computer applications developed to solve complex problem in a particular domain, at the level of extra-ordinary human intelligence and expertise.</p> <p>~Characteristics of Expert Systems</p> <ul style="list-style-type: none"> -High performance. -Understandable. -Reliable. - Highly responsive. <p>~Capabilities of Expert Systems The expert systems are capable of:</p> <ul style="list-style-type: none"> -Advising. -Instructing and assisting human in decision making.

- Demonstrating.
- Deriving a solution
- Diagnosing.
- Explaining.
- Interpreting input.
- Predicting results.
- Justifying the conclusion.
- Suggesting alternative options to a problem.

~In Capabilities of Expert Systems

They are incapable of:

- Substituting human decision makers.
- Possessing human capabilities.
- Producing accurate output for inadequate knowledge base.
- Refining their own knowledge.

Components of expert systems are:

- Knowledge base
- Interference engine
- User Interference

Development of expert systems:

1. Identify Problem Domain

- The problem must be suitable for an expert system to solve it. Find the experts in task domain for the ES project.
- Establish cost-effectiveness of the system.

2. Design the System

- Know and establish the degree of integration with the other systems and databases.
- Identify the ES Technology.
- Realize how the concepts can represent the domain knowledge best.

3. Develop the Prototype

From Knowledge Base: The knowledge engineer works to:

- Acquire domain knowledge from the expert.
- Represent it in the form of If-THEN-ELSE rules.

4. Test and Refine the Prototype

- The knowledge engineer uses sample cases to test the prototype for any deficiencies in performance.
- End users test the prototypes of the ES.

5. Develop and Complete the ES

- Test and ensure the interaction of the ES with all elements of its environment, including end users, databases, and other information systems.
- Document the ES project well.
- Train the user to use ES.

6. Maintain the System

- Keep the knowledge base up-to-date by regular review and update.
- Cater for new interfaces with other information systems, as those systems evolve.

Benefits of Expert Systems:

Availability: They are easily available due to mass production of software.

Less Production Cost: Production cost is reasonable. This makes them affordable.
 Speed: They offer great speed. They reduce the amount of work an individual puts in.
 Less Error Rate: Error rate is low as compared to human errors.
 Reducing Risk: They can work in the environment dangerous to humans.
 Steady response: They work steadily without getting motional, tensed or fatigued.

11. Explain forward chaining and backward chaining.

Inference engine

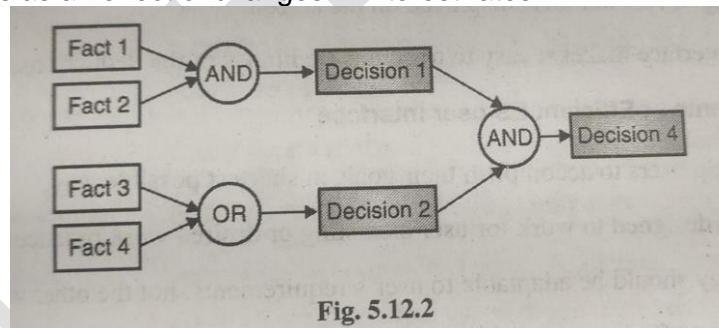
Use of efficient procedures and rules by the Inference Engine is essential in deducting a correct, flawless solution. In case of knowledge-based ES, the Inference Engine acquires and manipulates the knowledge from the knowledge base to arrive at a particular solution. In case of rule based ES, it:
 Applies rules repeatedly to the facts, which are obtained from earlier rule application.
 Adds new knowledge into the knowledge base if required.
 Resolves rules conflict when multiple rules are applicable to a particular case.

To recommend a solution, the Inference Engine uses the following strategies:

1. Forward Chaining
- 2 Backward Chaining

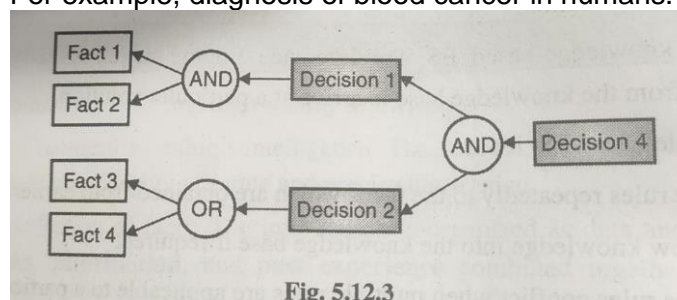
1. Forward Chaining:

It is a strategy of an expert system to answer the question, "What can happen next?" Here, the Inference Engine follows the chain of conditions and derivations and finally deduces the outcome. It considers all the facts and rules, and sorts them before concluding to a solution. This strategy is followed for working on conclusion, result, or effect. For example, prediction of share market status as an effect of changes in interest rates.



2.Backward Chaining:

With this strategy, an expert system finds out the answer to the question, "Why happened?" On the basis of what has already happened, the Inference Engine tries to find out which conditions could have happened in the past for this result. This strategy is followed for finding out cause or reason. For example, diagnosis of blood cancer in humans.



12.	What are the areas for expert system applications?
	<p>Interpretation Systems – Systems that infer situation descriptions from observations. This category includes surveillance, speech understanding, image analysis, signal interpretation and many kinds of intelligence analysis. An interpretation system explains observed data by assigning them symbolic meanings that describe the situation</p> <p>Prediction Systems – These systems include weather forecasting, demographic predictions, economic forecasting, traffic predictions, crop estimates, and military, marketing and financial forecasting</p> <p>Diagnostic Systems – These systems include medical, electronic, mechanical and software diagnoses. Diagnostic systems typically relate observed behavioural irregularities to underlying causes</p> <p>Design Systems – These systems develop configurations of objects that satisfy the constraints of the design problem. Such problems include circuit layout, building design and plant layout. Design systems construct descriptions of objects in various relationships with one another and verify that these configurations conform to stated constraints.</p> <p>Planning Systems – These systems specialize in planning problems such as automatic programming. They also deal with short and long term planning areas such as project management, routing, communications, product development, military applications and financial planning</p> <p>Monitoring Systems – These systems compare observations of system behaviour with standards that seem crucial for successful goal attainment. These crucial features correspond to potential flaws in the plan.</p> <p>Debugging Systems – These systems rely on planning, design and prediction capabilities for creating specifications or recommendations to correct a diagnosed problem</p> <p>Repair Systems – These systems develop and execute plans to administer a remedy for certain diagnosed problem. Such systems incorporate debugging, planning and execution capabilities.</p> <p>Instruction Systems – Systems that incorporate diagnosis and debugging subsystems that specifically address students need. These systems begin by constructing a hypothetical description of the student that interprets his behaviour. They then diagnosis his weakness and identify appropriate remedies to overcome the deficiencies. Finally they plan a tutorial interaction intended to deliver remedial knowledge to the student</p> <p>Control Systems – Systems that adaptively govern the overall behaviour of the system. To do this a control system must repeatedly interpret the current situation predict the future, diagnose the cause of anticipated problem, formulate a remedial plan and monitor its execution to ensure success.</p>

13.	What is knowledge engineering? Explain the process of knowledge engineering. / Write short note on knowledge engineering.
	<p>KNOWLEDGE ENGINEERING</p> <p>The collection of intensive activities encompassing acquisition of knowledge from human experts and conversion of this knowledge into repository is known as knowledge acquisition.</p> <p>Knowledge engineering requires cooperation and close communication between human experts and knowledge engineer to successfully codify and represent rules that human expert uses to solve problems within specific application domain.</p> <p>Knowledge possessed by human is often unstructured and not explicitly expressed. The major goal of knowledge engineering is to help experts articulate “how they do what they do” and to document this knowledge in a reusable form.</p> <p>Knowledge engineering deals with steps necessary to build expert system:</p> <ul style="list-style-type: none">• Knowledge acquisition: The process of getting knowledge from experts• Knowledge representation: Selecting the most appropriate structures to represent the knowledge.• Knowledge validation: Testing that the knowledge of ES is correct and complete.• Inference: Ability to reason.• Explanation: Ability to explain its advice.