GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)
Seshadri Rao Knowledge Village, Gudlavalleru – 521 356.

Department of Computer Science and Engineering

HANDOUT

on

DATA WAREHOUSING AND MINING

Vision

To be a Centre of Excellence in computer science and engineering education and training to meet the challenging needs of the industry and society.

Mission

- To impart quality education through well-designed curriculum in tune with the growing software needs of the industry.
- To be a Centre of Excellence in computer science and engineering education and training to meet the challenging needs of the industry and society.
- To serve our students by inculcating in them problem solving, leadership, teamwork skills and the value of commitment to quality, ethical behavior & respect for others.
- To foster industry-academia relationship for mutual benefit and growth

Program Educational Objectives

PEO1: Identify, analyze, formulate and solve Computer Science and Engineering problems both independently and in a team environment by using the appropriate modern tools.

PEO2: Manage software projects with significant technical, legal, ethical, social, environmental and economic considerations

PEO3: Demonstrate commitment and progress in lifelong learning, professional development, leadership and Communicate effectively with professional clients and the public.

HANDOUT ON DATA WAREHOUSING AND MINING

Class & Sem.	: III B.Tech – II Semester	Year : 2019-20
Branch	: CSE	Credits : 3

==

1. Brief History and Scope of the Subject

The term "Data Mining" was only introduced in the 1990s. Data mining is part of the knowledge discovery process that offers a new way to look at data. Data mining consists of the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. It uses machine learning, statistical and visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans. Data mining is then the process of discovering meaningful new correlations, patterns and trends by sifting through vast amounts of data using statistical and mathematical techniques.

As Fortune 500 organizations continue to amass substantial quantities of information into their respective databases, data mining can offer the opportunity to learn from this data. Furthermore, current trends indicate that more companies implementing Enterprise Resource Planning systems or contracting with ASP vendors could further benefit in using data mining techniques. Integrating a data mining technique alongside these two added value services can proof to be an optimum solution in understanding a company's data.

2. Pre-Requisites

- Database Management Systems, Basics of Probability and Statistics

3. Course Objectives:

- To introduce the concepts of Data warehousing and Data mining.
- To familiarize with the concepts of association rule mining, classification, clustering techniques and algorithms.

4. Course Outcomes:

CO1: Outline different types of databases used in data mining

CO2: Apply pre-processing methods on raw data to make it ready for mining.

CO3: Illustrate the major concepts and operations of multi dimensional data models.

CO4: Analyze the performance of association rules mining algorithms for finding frequent item sets from the large databases

CO5: Simplify the data classification procedure by selecting appropriate classification methods / algorithms

CO6: Classify various clustering methods and algorithms on data sets to create appropriate clusters.

5. Program Outcomes:

Computer Science and Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including

prediction and modelling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

6. Mapping of Course Outcomes with Program Outcomes:

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	12
CO1	3	2	3	2								
CO2	3	2	2		3	2						
CO3			3	2								
CO4	2	2	2		2	3						
CO5		2	2		3							
CO6	2	2	2	3	3							

3- High Level Mapping **2-** Medium Level Mapping **1-**Low Level Mapping

7. Prescribed Text Books

1. Jiawei Han & Micheline Kamber, & Jian pei,“Data Mining Concepts and Techniques”, 3rd edition, Morgan Kaufmann Publisher an imprint of Elsevier.

8. Reference Text Books

a. Pang-Ning Tan, Michael Steinbach, Vpin Kumar “Introduction to Data Mining”, 1st edition, Pearson.

b. Margaret H Dunham, “Data Mining Introductory and Advanced Topics”, 1st edition, Pearson Education

9. URLs and Other E-Learning Resources

a. http://www.cs.sfu.ca/~han/dmbook
b. http://db.cs.sfu.ca/
c. http://www.cs.sfu.ca/~han

10. Digital Learning Materials:

- http://192.168.0.49/videos/videosListing/270#

11. Lecture Schedule / Lesson Plan

Topic	No. of Periods
UNIT - I: INTRODUCTION	
Motivation and importance of data mining	2
Types of data to be mined: Relational database, datawarehouses, transactional databases, advanced database systems	4
Data Mining Functionalities	2
	8
UNIT - II: DATA PRE-PROCESSING	
Major tasks in data pre-processing	1

Data cleaning: Missing values, Noisy Data	2
Data reduction: Overview of data reduction strategies, Principal components analysis Attribute subset selection, histograms, sampling	4
Data Transformation: Data transformation strategies overview, data transformation by normalization	3
	10
UNIT - III: DATA WAREHOUSING AND ONLINE ANALYTICAL PROCESSING	
Data warehouse: Basic concepts, OLAP vs OLTP	2
Data warehouse: A multi-tired architecture	1
Data warehouse modeling : Data cube and OLAP	2
Data cube: A multidimensional data model, star, snowflake and fact constellation schemas for multidimensional data models	3
The role of concept hierarchies	1
Typical OLAP operations	1
	10
UNIT - IV: MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS	
Basic concepts, Frequent item sets, closed item sets and association rules	2
Frequent item set mining methods: Apriori Algorithm, generations, association rules from frequent item sets	3
A Pattern-Growth approach for mining frequent item sets	2
	7
UNIT - V: CLASSIFICATION	
Basic concepts, What is classification, general approach to classification	2
Decision Tree Induction	2
Attribute selection measures : Information gain	3
Bayes classification methods: Bayes' theorem	2

Naïve Bayesian classification	2
	11
UNIT - VI: CLUSTER ANALYSIS	
Introduction, Overview of basic clustering methods	2
Partitioning methods: k-means, k-medoids	3
Hierarchical methods: Agglomerative versus divisive hierarchical clustering	3
Density based method: DBSCAN	2
	10
Total No.of Periods:	**56**

12. Seminar Topics:

In order to enhance the understanding capability and to prepare the student to face the interviews and audience, to enhance the communication skills and to eliminate stage fear, seminars and group discussions are conducted.

- Data Warehouse and OLAP
- Concept Hierarchy Generation
- Bayesian Classification
- Density-Based Methods

UNIT – I

Objective:

- To gain knowledge on Data mining

Syllabus:

Unit-I:

Motivation and importance of data mining, types of data to be mined: Relational databases, data warehouses, transactional databases, advanced database systems, data mining functionalities.

Learning Outcomes:

At the end of the unit, students will be able to:

1. Understand functionalities of Data Mining
2. Identify and study different databases to implement data mining systems.

Learning Material

Introduction

1.1 Motivation and importance of data mining

- Motivation and Importance of Data Mining in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.
- The information and knowledge gained can be used for applications ranging from business management, production control and market analysis to engineering design and science exploration.

Definition : Data mining refers to extracting or "mining" knowledge from large amounts of data.

(or)

It is the process of automatically discovering useful information in large data repositories.

- Data mining should have been more appropriately named “knowledge mining from data,” in short it is “Knowledge mining,” many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

Different views of Data mining

1) Data mining as simply an essential step in the process of knowledge discovery.

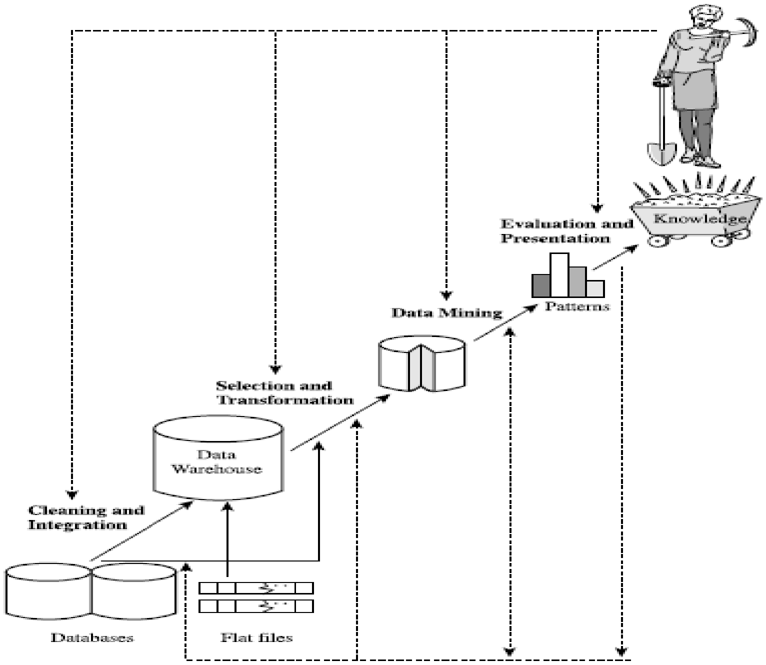


Fig 1: Data mining as a step in the process of knowledge discovery

Above Figure consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques
are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

according to this view, data mining is only one step in the entire process, and it is essential one because it uncovers hidden patterns for evaluation.

2) Other view of Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.

Based on this view, the architecture of a typical data mining system may have the following major components

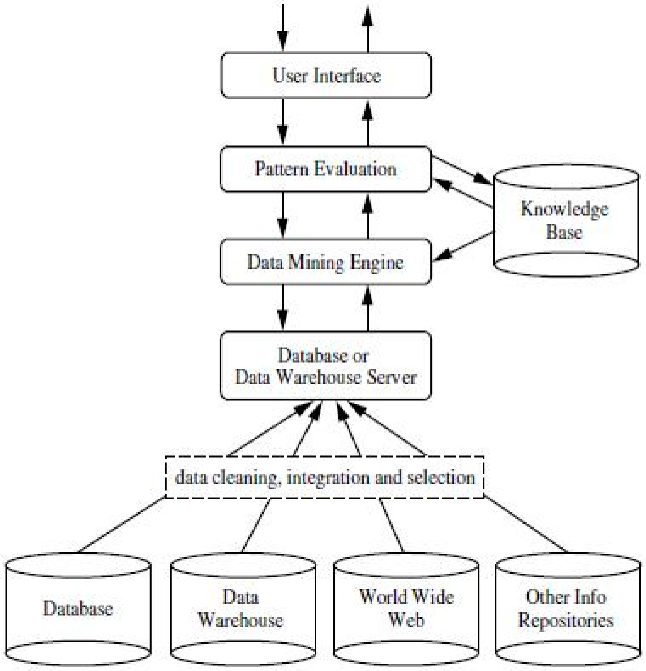


Fig 2 : Architecture of a typical data mining system.

- **Database, data warehouse, WorldWideWeb, or other information repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories.
- **Data cleaning and data integration** techniques may be performed on the data.
- **Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.

  Ex: concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

- **Data mining engine:** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.
- **Pattern evaluation module:** This component typically employs interestingness measures. The pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used
- **Graphical User interface:** This module communicates between users and the data mining system.

  Allowing the user

  - To interact with the system by specifying a data mining query or task.
  - Providing information to help focus the search.
  - Performing exploratory data mining based on the intermediate data mining results.
  - Allows the user to browse database and data warehouse schemas or data structures.
  - Evaluate mined patterns, and visualize the patterns in different forms.

- Data mining involves an integration of techniques from multiple disciplines such as
  - Database technology
  - Statistics
  - Machine learning
  - High-performance computing
  - Pattern recognition
  - Neural networks
  - Data visualization

- Information retrieval
- Image and signal processing
- Spatial or Temporal data analysis

❖ By performing data mining, interesting knowledge, high-level information can be extracted from databases and viewed from different angles.

❖ The discovered knowledge can be applied to decision making, process control, information management and query processing.

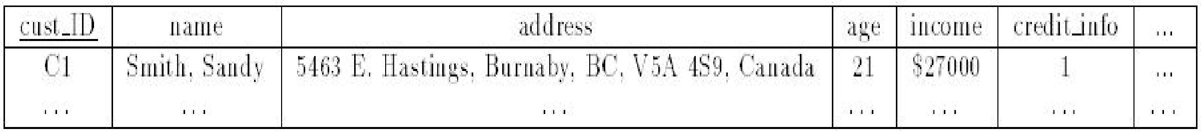
❖ The data mining considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry.

1.2 Data mining should be applicable to any kind of data

1) Relational databases

- A Relational database is defined as the collection of data organized in tables with rows and columns.
- Physical schema in Relational databases is a schema which defines the structure of tables.
- Logical schema in Relational databases is a schema which defines the relationship among tables.
- Standard API of relational database is SQL.
- **Application**: Data Mining, ROLAP model, etc.

customer



cust_ID	name	address	age	income	credit_info	...
C1	Smith, Sandy	5463 E. Hastings, Burnaby, BC, V5A 4S9, Canada	21	$27000	1	...
...	...	...	...	...	...	...

STUDENT

ROLL_NO	NAME	ADDRESS	PHONE	AGE
1	RAM	DELHI	9455123451	18
2	RAMESH	GURGAON	9652431543	18
3	SUJIT	ROHTAK	9156253131	20
4	SURESH	DELHI		18

2) DWH

- A data warehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
- There are three types of data warehouse: **Enterprise** data warehouse, **Data Mart** and **Virtual** Warehouse.
- Two approaches can be used to update data in Data Warehouse: **Query-driven** Approach and **Update-driven** Approach.
- **Application**: Business decision making, Data mining, etc.

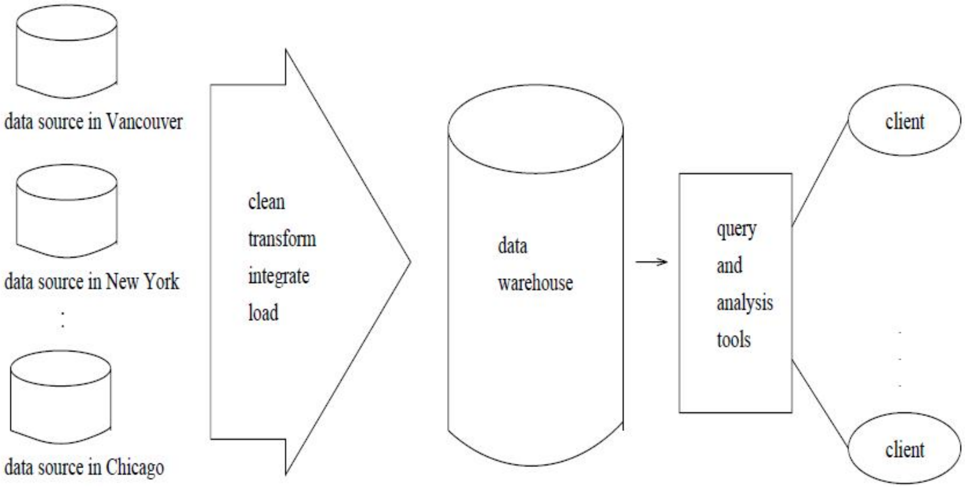
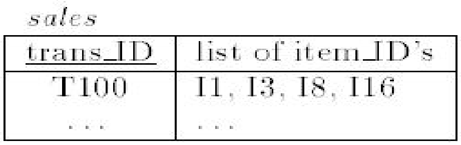


Figure: typical architecture of a data warehouse for AllElectronics.

3) Transactional databases

- Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
- This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
- Highly flexible system where users can modify information without changing any sensitive information.
- Follows ACID property of DBMS.
- **Application**: Banking, Distributed systems, Object databases, etc.



sales

trans_ID	list of item_ID's
T100	I1, I3, I8, I16
...	...

Figure: transactional database for sales at AllElectronics

4) Advanced databases

- Object oriented
- Object relational
- Application oriented databases
  - Spatial
  - Temporal
  - Time-Series
  - Text
  - Multimedia databases

Multimedia Databases

- Multimedia databases consists audio, video, images and text media.
- They can be stored on Object-Oriented Databases.
- They are used to store complex information in a pre-specified formats.

- **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.

Spatial Database

- Store geographical information.
- Stores data in the form of coordinates, topology, lines, polygons, etc.
- **Application**: Maps, Global positioning, etc.

Time-series Databases

- Time series databases contains stock exchange data and user logged activities.
- Handles array of numbers indexed by time, date, etc.
- It requires real-time analysis.
- **Application**: eXtremeDB, Graphite, InfluxDB, etc.

1.3 Data Mining Functionalities—What Kinds of Patterns Can Be Mined?

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- Data mining tasks can be classified into two categories:
  1) Descriptive
  2) Predictive
     - Descriptive tasks derive patterns that summarizes the underlying relationship in the data. Ex: correlations, trends, clusters, trajectories and anomalies. These are in explanatory in nature.
     - Predictive tasks perform inference on the current data to make predictions. i.e predict the value of a particular attribute based on the values of other attributes. ex: classification, regression.

Data mining functionalities, and the kinds of patterns they can discover, are described below:

1. Characterization & Discrimination
2. Association analysis
3. Classification
4. Evolution analysis
5. Clustering
6. Outlier analysis.

1.3.1. Concept/Class Description: Characterization and Discrimination

Data can be associated with **classes or concepts**.

Ex: In the AllElectronics store,

classes of items for sale include computers and printers

concepts of customers include bigSpenders and budgetSpenders.

The summarized descriptions of class or a concept are very much useful. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived via (1) data characterization 2) data discrimination, (3) both data characterization and discrimination

Data characterization is a summarization of the general characteristics or features of data.

Ex: study the characteristics of software products whose sales increased by 10% in the last year.

- Methods used for this are statistical measures, plots and OLAP operations.
- The output of data characterization can be presented in various forms.

  Ex:pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables,

Data discrimination is comparison of the target class(the class under study) with one or a set of comparative classes (called the contrasting classes).

Ex: the user may like to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period

- Methods used and output presentation is same as characterization although discrimination descriptions should include comparative measures that help distinguish between the target and contrasting classes

1.3.2. Association Analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. This analysis is widely used for market basket or transaction data analysis.

Association rules are of the form **X => Y,** is interpreted as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y".

Ex: Marketing manager of AllElectronics, would like to determine which items are frequently purchased together within the same transactions. An example of such a rule, mined from the AllElectronics transactional database

buys(X, "computer") => buys(X, "software") [support = 1%; confidence = 50%]

where X is a variable representing a customer. A confidence of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well, and 1% of all of the transactions contain both computer and software were purchased together

1.3.3. Classification and Prediction

Classification is the process of finding a model that describes and distinguishes data classes or concepts. To predict the class of objects whose class label is un known. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks

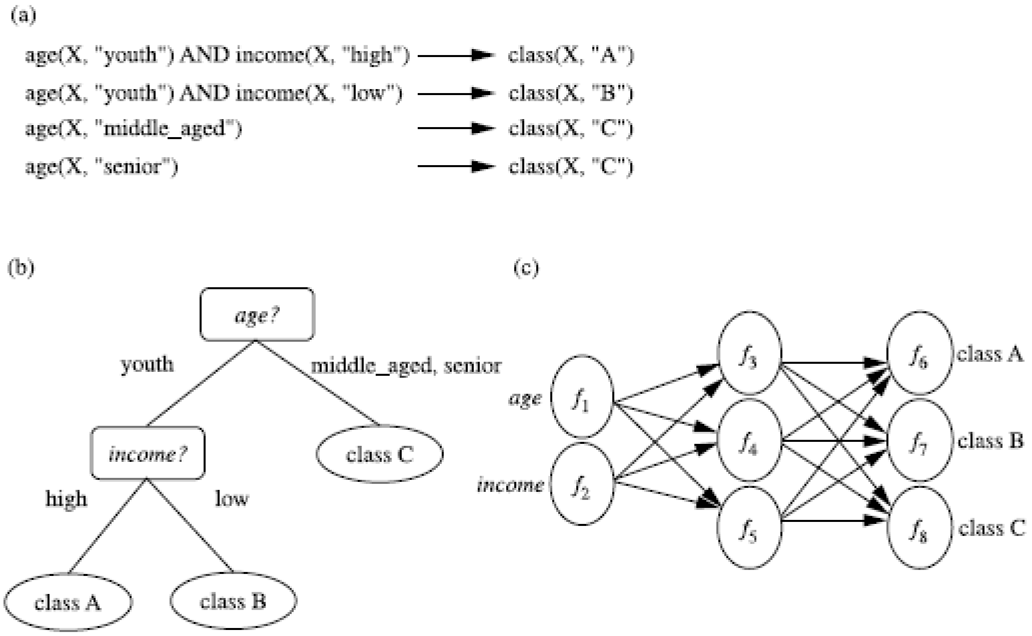

(a)
age(X, "youth") AND income(X, "high") → class(X, "A")
age(X, "youth") AND income(X, "low") → class(X, "B")
age(X, "middle_aged") → class(X, "C")
age(X, "senior") → class(X, "C")
(b)
age?
youth
middle_aged, senior
income?
class C
high
low
class A
class B
(c)
age
income
f_1
f_2
f_3
f_4
f_5
f_6
f_7
f_8
class A
class B
class C

Fig: classification model can be represented in various forms, such as (a) IF-THEN rules,(b) a decision tree, or a (c) neural network.

Ex: AllElectronics, items are classified into 3 classes good response, mild response and no response. based on the descriptive features of the items based ons price, brand, place made, type, and category.

Predict missing or unavailable data values are referred as Prediction.

1.3.4.Evolution Analysis

It describes and models regularities or trends for objects whose behavior changes over time, this may include characterization, discrimination, association and correlation analysis, classification, prediction, clustering.

Ex: Stock market data analysis to predict the future trends using previous years data for decision making regarding stock investments.

1.3.5.Cluster Analysis

Cluster is a group of similar data points or objects for analysis. The objects within a cluster have high similarities in comparison to one another but are very dissimilar to objects in other clusters.

Ex: Cluster AllElectronics customer data with respect to customer locations in a city. These clusters may represent individual target groups for marketing.

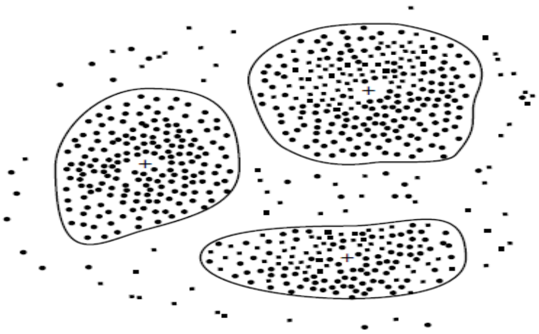


Fig : 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster “center” is marked with a “+”.

6. Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers(noise in the data) outliers may be detected using statistical tests

Ex : Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

UNIT-I
Assignment-Cum-Tutorial Questions
SECTION-A

Objective Questions

1. _______is the process of discovering interesting patterns and knowledge from large amounts of data.

2. The full form of KDD is________

3. Goal of data mining includes which of the following [ ]

   A. To explain some observed event or condition

   B. To confirm that data exists

   C. To analyze data from expected relationships

   D. To create a new data warehouse

4. The Synonym for data mining is [ ]

A Data warehouse B) Knowledge Discovery from Data C) ETL D) OLAP

5. Data mining tasks are classified in to _____________ and _____________.

6. Match the Following: [ ]

a) Data Cleaning.	i) Multiple data sources may be combined
b) Data Transformation	ii) Remove noise and inconsistent data
c) Data Selection	iii) Data transformed into forms appropriate for mining
d) Data Integration	iv) Relevent data is retrived from database for analysis.

A. i,ii,iii,iv B. i,iii,iv,ii C. ii,iii,iv,i D. iv,ii,iii,i

7. Data mining helps in __________. [ ]

A. inventory management. C.sales promotion strategies

B. marketing strategies. D.All of the above

8. Which of the following is not a data mining functionality? [ ]

A. Characterization and Discrimination C. Classification and regression

B. Selection and interpretation D. Clustering and Analysis

6. Extreme values that occur infrequently are called as _________. [ ]

A. outliers. B. rare values. C. dimensionality reduction. D. All

7.Grouping of similar objects is known as ____________

8. Support and Confidence are used as a measures for Association Rule Mining. [T/F]

9___________ is a summarization of the general characteristics or features of a target class of data. [ ]

A. Data Characterization B. Data Classification
C. Data discrimination D. Data selection

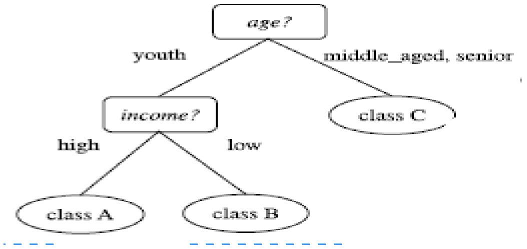
10Match the following Issues: [ ]

a) Mining Methodology. i) Efficiency and Salability
b) User Interaction ii) Handling of relational and complex types of Data
c) Diverse Datatypes iii) Interactive Mining of Knowledge at multiple levels of abstration
d) Performance iv) Mining different kinds of knowledge in databases.

A. i,ii,iii,iv B. i,iii,iv,ii C. ii,iii,iv,i D. iv,iii,ii,i

11. _____________ is the process of finding a model that describes and distinguishes data classes or concepts.

12.The Following diagram represents _________ Model. [ ]



A. Classification B. Cluster C. Evolution D. Association

13. __________ Analysis can be used for unlabeled dataset.

14. What mining task characterizes properties of the data in a target data set?

A) Predictive B) Descriptive C) Both D) None of the above [ ]

SECTION-B

SUBJECTIVE QUESTIONS

1. Write briefly about motivation of challenges for data mining.
2. Define Data Mining. Explain the steps to discover knowledge.
3. Write few disciplines where Data mining is applied.
4. Explain various kinds of databases.
5. What are advanced data base systems?
6. Differentiate operational databases and data warehousing.
7. What are Data mining Functionalities? Explain.