INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

INFORMATION TECHNOLOGY

DEFINITIONS AND TERMINOLOGY

Course Name	:	**DATA WAREHOUSING AND DATA MINING**
Course Code	:	**AIT006**
Program	:	**B.Tech**
Semester	:	**VI**
Branch	:	**CSE / IT**
Section	:	**A,B,C,D**
Academic Year	:	**2019– 2020**
Course Faculty	:	**Dr.M Madhubala,Professor and HOD** **Dr. D.Kishore Babu, Associate. Professor** **Mr.Ch Suresh Kumar Raju, Assistant. Professor** **Mr.A Praveen, Assistant. Professor** **Ms.S Swarajya Lakshmi, Assistant. Professor** **Ms.G Geetha Yadav, Assistant. Professor**

COURSE OBJECTIVES (COs):

The course should enable the students to:	
I	Identifying necessity of Data Mining and Data Warehousing for the society..
II	Familiar with the process of data analysis, identifying the problems, and choosing the relevant models and algorithms to apply
III	Develop skill in selecting the appropriate data mining algorithm for solving practical problems.

IV	Develop ability to design various algorithms based on data mining tools
V	Create further interest in research and design of new Data Mining techniques and concepts.

DEFINITIONS AND TERMINOLOGYQUESTION BANK

S No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
UNIT - I						
1	What is Data?	Data is raw, unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized.	Understand	CO 1	CLO04	AIT006.04
2	What is Information?	When data is processed, organized, structured or presented in a given context so as to make it useful, it is called information.	Understand	CO 1	CLO 03	AIT006.03
3	What is Database?	A database is a collection of information that is organized so that it can be easily accessed, managed and updated. Data is organized into rows, columns and tables, and it is indexed to make it easier to find relevant information.	Understand	CO 1	CLO08	AIT006.08
4	Define Data warehouse?	A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.	Understand	CO 1	CLO04	AIT006.04
5	Define data mart.	Data mart can be defined as the subset of data warehouse of an organization which is limited to a specific business unit or group of users.		CO 1	CLO03	AIT006.03
6	Define enterprise warehouse	An enterprise data warehouse is a unified database that holds all the business information an organization and makes it accessible all across the company.		CO 1	CLO04	AIT006.04
7	Define Virtual warehouse	A virtual warehouse is another term for a data warehouse. A data warehouses a computing tool designed to simplify decision-making in business		CO 1	CLO03	AIT006.03

		management. It collects and displays business data relating to a specific moment in time, creating a snapshot of the condition of the business at that moment				
8	Define data repository	Data Repository is a logical (and sometimes physical) partitioning of data where multiple databases which apply to specific applications or sets of applications reside.	Understand	CO 1	CLO04	AIT006.04
9	Define meta data	Metadata is data that describes other data. Meta is a prefix that in most information technology usages means "an underlying definition or description	Understand	CO 1	CLO03	AIT006.03
10	What is time sharing	Time-sharing is a technique which enables many people, located at various terminals, to use a particular computer system at the same time. Time-sharing or multitasking is a logical extension of multiprogramming. Processor's time which is shared among multiple users simultaneously is termed as time-sharing.	Remember	CO 1	CLO03	AIT006.03
11	Define OLAP cube	An OLAP cube is a multidimensional database that is optimized for data warehouse and online analytical processing (OLAP) applications. An OLAP cube is a method of storing data in a multidimensional form, generally for reporting purposes.	Understand	CO 1	CLO 03	AIT006.03
12	Difference between Database and Data Warehouse?	A data warehouse is built to store large quantities of historical data and enable fast, complex queries across all the data, typically using Online Analytical Processing (OLAP). A database was built to store current transactions and enable fast access to specific transactions for ongoing business processes, known as Online Transaction Processing (OLTP).	Remember	CO 1	CLO03	AIT006.03
13	Name the OLAP operations.	Roll-up Drill-down Slice and dice	Understand	CO 1	CLO03	AIT006.03

14	Define Data warehouse?	A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process	Remember	CO 1	CLO03	AIT006.03
15	What is olap?	OLAP (online analytical processing) is a computing method that enables users to easily and selectively extract and query data in order to analyze itfrom different points of view.	Understand	CO 1	CLO05	AIT006.05
16	Differences Between Star And Snowflake Schemas?	Star schema - all dimensions will be linked directly with a fat table. Snow schema - dimensions maybe interlinked or may have one-to-many relationship with other tables.	Remember	CO 1	CLO 01	AIT006.01
17	What Is Etl?	ETL stands for extraction, transformation and loading.ETL provide developers with an interface for designing source-to-target mappings, transformation and job control parameter.	Understand	CO 1	CLO03	AIT006.03
18	What is aggregation?	Data aggregation is any process in which information is gathered and expressed in a summary form, for purposes such as statistical analysis. A common aggregation purpose is to get more information about particular groups based on specific variables such as age, profession, or income.	Understand	CO 1	CLO 10	AIT006.10
19	Elaborate MOLAP?	Multi dimensional online analytical processing	Remember	CO 1	CLO 14	AIT006.14
20	Define concept hierarchy?	A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.	Understand	CO 1	CLO04	AIT006.04
21	Differentiate between slice and dice?	The slice operation selects one particular dimension from a given cube and provides a new sub-cube and Dice selects two or more dimensions from a given cube and provides a new sub-cube	Understand	CO 1	CLO05	AIT006.05
22	What is OLTP?	OLTP (online transaction processing) is a class of software programs capable of supporting transaction-oriented applications on the Internet	Understand	CO 1	CLO 09	AIT006.09
23	Define star schema?	Star schema is the simplest form of a dimensional model, in which data is organized into facts and dimensions.	Understand	CO 1	CLO 08	AIT006.08

24	Define fact table?	A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often normalized	Understand	CO 1	CLO 08	AIT006.08
25	Define dimension table?	A dimension table is a table in a star schema of a data warehouse. A dimension table stores attributes, or dimensions, that describe the objects in a fact table.	Understand	CO 1	CLO 02	AIT006.02
26	define fact constellation schema?	Fact constellation is a collection of multiple tables sharing dimension tables viewed as a collection of stars. It can be seen as an extension of the schema. A fact constellation schema has multiple fact tables. It is also known as galaxy schema	Understand	CO 1	CLO03	AIT006.04
27	What Is Ods?	ODS means Operational Data Store. A collection of operation or bases data that is extracted from operation databases and standardized, cleansed, consolidated, transformed, and loaded into an enterprise data architecture.	Understand	CO 1	CLO04	AIT006.04
28	Elaborate ROLAP?	Relational online analytical Processing	Remember	CO 1	CLO04	AIT006.04
29	Describe about Drill-down Operation?	Drill-down is the reverse operation of roll-up. It is performed by either of the following ways By stepping down a concept hierarchy for a dimension By introducing a new dimension	Understand	CO 1	CLO15	AIT006.15
30	Describe about update-driven approach	In update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis	Understand	CO 1	CLO03	AIT006.03
31	Define snowflake schema?	Snowflake is a form of dimensional modeling in which dimensions are stored in multiple related dimension tables. A snowflake schema is a variation of the star schema.	Understand	CO 1	CLO 13	AIT006.13
UNIT – II						
1	What is Data Mining?	Data Mining is the process of Extraction of implicit knowledge from multiple heterogeneous data sources in the form of patterns.	Understand	CO 2	CLO 07	AIT006.07

2	Describe Heterogeneous data sources.	Heterogeneous data sources includes Relational, Object-oriented, Object based, flat files, www, image, audio and text data sources.	Remember	CO 2	CLO 08	AIT006.08
3	Elaborate KDD	Knowledge Discovery in Databases.	Remember	CO 2	CLO 08	AIT006.08
4	What is data cleaning?	Data Cleaning is the process of identifying and removing the noise and inconsistent data.	Understand	CO 2	CLO 04	AIT006.04
5	What is Data Integration?	Data Integration is the process of combining the data from multiple data sources.	Understand	CO 2	CLO 08	AIT006.08
6	What is data Selection?	Data Selection is the process of retrieving analysis task relevant data from the database	Understand	CO 2	CLO 08	AIT006.08
7	List the different measures to evaluate the pattern / rules.	Objective and subjective interestingness measures: Objective: based on statistics and structures of patterns, e.g., support, confidence, etc. Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, action ability, etc.	Remember	CO 2	CLO 10	AIT006.10
8	How to perform Data Transformation?	Data Transformation is the process of transforming or consolidating into forms. Which are appropriate for mining by performing summary or aggregation operations.	Understand	CO 2	CLO 09	AIT006.09
9	Describe Concept / Class Description.	Concept / Class Description of the data is performed with Characterization and discrimination of the data. Which need to perform Generalization, summarization, and finding contrast data characteristics	Understand	CO 2	CLO 08	AIT006.08
10	Differentiate classification and prediction?	Classification is Finding models (functions) that describe and distinguish classes or concepts for future prediction. Prediction is Predict some unknown or missing numerical values.	Understand	CO 2	CLO 09	AIT006.09

11	Differentiate Between Data Mining And Data Warehousing?	Data warehousing is merely extracting data from different sources, cleaning the data and storing it in the warehouse. Whereas data mining aims to examine or explore the data using queries.	Understand	CO 2	CLO 08	AIT006.08
12	Difference between descriptive and predictive data mining?	Descriptive data mining, which describes data in a concise and summative manner .Predictive data mining, which analyzes data in order to construct one or a set of models and attempts to predict the behavior of new datasets	Understand	CO 2	CLO 10	AIT006.10
13	List out Different Stages Of "data Mining"?	Exploration Model building and validation Deployment.	Remember	CO 2	CLO 08	AIT006.08
14	Define legacy data base?	A legacy data source is any file, database, or software asset (such as a web service or business application) that supplies or produces data and that has already been deployed.	Understand	CO 2	CLO 07	AIT006.07
15	Define Descriptive Model	Descriptive modeling is a mathematical process that describes real-world events and the relationships between factors responsible for them.	Understand	CO 2	CLO 09	AIT006.09
16	Define Program counter	Program counter defined as the counter indicates the address of the next instruction to be executed for this process	Remember	CO 2	CLO 09	AIT006.09
17	Definition of binning?	Binning is a way to group a number of more or less continuous values into a smaller number of "bins".	Understand	CO 2	CLO 08	AIT006.08
18	Difference between Discrete And Continuous Data In Data Mining World?	Discrete data can be considered as defined or finite data. E.g. Mobile numbers, gender. Continuous data can be considered as data which hangs continuously and in an ordered fashion. E.g. age.	Remember	CO 2	CLO 08	AIT006.08
19	How Does The Data Mining And Data Warehousing Work Together?	Data warehousing can be used for analyzing the business needs by storing data in a meaningful form. Using Data mining, one can forecast the business needs. Data warehouse can act as a source of this forecasting.	Remember	CO 2	CLO 04	AIT006.04

20	How do you clean the data?	Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.	Understand	CO 2	CLO 10	AIT006.10
21	What is required technological drivers in data mining?	Database size: Basically, as for maintaining and processing the huge amount of data, we need powerful systems.	Remember	CO 2	CLO 08	AIT006.08
22	Define shared-memory	The shared-memory method requires communicating processes to share some variables.	Remember	CO 2	CLO 07	AIT006.07
23	List the issues / challenges in Data mining system.	1. Mining methodology and user interaction 2. Performance and scalability 3. Diversity of data types Applications and social impacts	Remember	CO 2	CLO 10	AIT006.10
24	List the DM task primitives	1. Set of task-relevant data to be mined 2. Kind of knowledge to be mined	Remember	CO 2	CLO 08	AIT006.08
25	List the issues / challenges in Data mining system	1. Mining methodology and user interaction 2. Performance and scalability 3. Diversity of data types Applications and social impacts	Remember	CO 2	CLO 08	AIT006.08
26	List out The Issues Regarding Classification And Prediction?	Data cleaning o Relevance analysis o Data transformation o Comparing classification methods o Predictive accuracy o Speed o Robustness o Scalability Interpretability	Remember	CO 2	CLO 08	AIT006.08
27	Write the strategies for data reduction	1. Data cube aggregation 2. Attribute subset selection 3. Dimensionality reduction 4. Numerosity reduction 5. Discretization and concept hierarchy generation	Remember	CO 2	CLO 10	AIT006.10
28	How do you clean the data?	smooth out noise while identifying outliers, and correct inconsistencies in the data	Remember	CO 2	CLO 08	AIT006.08
29	List the issues / challenges in Data mining system	1. Mining methodology and user interaction 2. Performance and scalability 3. Diversity of data types Applications and social impacts	Remember	CO 2	CLO 08	AIT006.08
30	List the DM task primitives	1. Set of task-relevant data to be mined 2. Kind of knowledge to be mined	Remember	CO 2	CLO 10	AIT006.10
UNIT – III						
1	What are frequent patterns?	Frequent Pattern is a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data setstorage. In computer architecture, frames are analogous to logical address space pages.	Remember	CO 3	CLO 13	AIT006.13

2	List the Applications of frequent pattern analysis?	Basket data analysis cross-marketing catalog design sale campaign analysis Web log (click stream) analysis and DNA sequence analysis	Remember	CO 3	CLO 14	AIT006.14
3	What is an Association rule mining?	Association rule mining is a procedure which is meant to find frequent patterns, correlations ,associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories..	Remember	CO 3	CLO 11	AIT006.11
4	State the general form of Association rule	Body => Head [support, confidence] Support and Confidence are measures to define strength of the rule.	Remember	CO 3	CLO 13	AIT006.13
5	Describe buys(x, “milk”) => buys(x, “bread”) [0.5%, 60%]	The rule says that 20% customers buy milk and bread together, and those who buy milk also buy bread 60% of the time.	Remember	CO 3	CLO 13	AIT006.13
6	List the different types of Association Rules	Boolean vs. Quantitative associations • Single dimension vs. Multiple dimensional associations Single level vs. multiple-level analysis	Remember	CO 3	CLO11	AIT006.11
7	State an example of Boolean vs. Quantitative association rule	buys(x, “SQLServer”) ^ buys(x, “DMBook”) => buys(x, “DBMiner”) [0.2%, 60%] age(x, “30..39”) ^ income(x, “42..48K”) => buys(x, “PC”) [1%, 75%]	Remember	CO 3	CLO 14	AIT006.14
8	Give an example of Single dimension vs. Multiple dimensional association rule	Single dimension Plays(cricket) => Plays(tennies) Multiple dimensions age(X,”20..25”) Λ income(X,”30K..41K”)buys (X,”Laptop Computer”)	Remember	CO 3	CLO 13	AIT006.13
9	Describe multilevel association rules	Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.”	Remember	CO 3	CLO11	AIT006.11
10	Define Confidence	The confidence of a rule, X → Y, is the percentage of transactions in T that contain X also contain Y, and can be seen as an estimate of the conditional probability, Pr(Y | X). It is computed as follows: Confidence(X→ Y) = (X∪Y)count / Xcount	Remember	CO 3	CLO14	AIT006.14

11	State Apriori principle	Apriori principle states that, If an itemset is frequent, then all of its subsets must also be frequent	Remember	CO 3	CLO 13	AIT006.13
12	List the computational challenges in Apriori Algorithm	• Huge number of candidates • Multiple scans of transaction database Tedious workload of support counting for candidates	Remember	CO 3	CLO11	AIT006.11
13	State the drawback in Apriori method	Huge number of candidates • Multiple scans of transaction database Tedious workload of support counting for candidates	Remember	CO 3	CLO 11	AIT006.11
14	List the methods to improve Apriori’s efficiency	Hash-based itemset counting • Transaction reduction • Partitioning • Sampling Dynamic itemset counting	Remember	CO 3	CLO 14	AIT006.14
15	What is an itemset?	A set of one or more items in a transaction	Remember	CO 3	CLO 11	AIT006.11
16	Define support?	Support: It is one of the measure of interestingness. This tells about usefulness and certainty of rules	Remember	CO 3	CLO11	AIT006.11
17	Define confidence	Confidence: A confidence of 60% means that 60% of the customers who purchased a milk and bread also bought butter	Remember	CO 3	CLO 13	AIT006.13
18	Definition of closed itemset ?	It is a frequent itemset that is both closed and its support is greater than or equal to minsup. An itemset is closed in a data set if there exists no superset that has the same support count as this original itemset.	Remember	CO 3	CLO 14	AIT006.14

S No	QUESTION	ANSWER	Blooms Level	CO	CLO	CLO Code
19	What are Iceberg queries?	Iceberg queries are a special case of SQL queries involving GROUP BY and HAVING clauses, wherein the answer set is small relative to the database size. Iceberg queries have been recently identified as important queries for many applications.	Remember	CO 3	CLO 14	AIT006.14
20	State the absolute measure of frequent patterns?	Support is the absolute measure of frequent measures. support count of X: Frequency or occurrence of an itemset X.	Remember	CO 3	CLO 11	AIT006.11
21	State the relative measure of frequent patterns?	(relative) support, s, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)	Remember	CO 3	CLO 11	AIT006.11
22	State itemset X is frequent or not?	An itemset X is frequent if X's support is no less than a minsup threshold value	Remember	CO 3	CLO 13	AIT006.13
23	What is FP-tree	Frequent Pattern tree uses a divide and conquer method for finding frequent itemsets..	Remember	CO 3	CLO11	AIT006.11
24	difference between Boolean association rule and quantitative Association rule.?	If a rule involves associations between the presence or absence of items, it is a Boolean association rule. Quantitative association rules involve numeric attributes that have an implicit ordering among values (e.g., age). If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule	Remember	CO 3	CLO 11	AIT006.11
25	What are the Meta rules are useful in constraint based association mining?	Meta rules may be based on the analyst's experience, expectations, or intuition regarding the data or may be automatically generated based on the database schema	Remember	CO 3	CLO 14	AIT006.14
26	List the techniques to improve the efficiency of Apriori algorithm	Hash based technique Transaction Reduction Portioning Sampling Dynamic item counting	Remember	CO 3	CLO 13	AIT006.13
27	Define association rule?	Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps: A minimum support threshold is applied to find all frequent itemsets in a database.	Remember	CO 3	CLO14	AIT006.14

28	Definition of Maximal frequent item set	The definition says that an item set is maximal frequent if none of its immediate supersets is frequent	Remember	CO 3	CLO 13	AIT006.13
29	Define Data	Unorganized facts that need to be processed. Data can be something simple and seemingly random and useless until it is organized	Remember	CO 3	CLO11	AIT006.11
30	Define Data Mining	Extracting knowledge from large amount of data	Remember	CO 3	CLO11	AIT006.11
UNIT - IV						
1	What is prediction?	Prediction models continuous-valued functions, i.e., predicts unknown or missing values	Remember	CO 4	CLO 15	AIT006.15
2	List the Major issues related to Classification	Data cleaning, Relevance analysis ,Data transformation	Remember	CO 4	CLO 15	AIT006.15
3	State the major steps in Classification	A two step process. Which includes: Model construction: describing a set of predetermined classes Model usage: for classifying future or unknown objects	Remember	CO 4	CLO 15	AIT006.15
4	Define supervised learning	Supervised learning (Classification): The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations and new data is classified based on the training set	Remember	CO 4	CLO 12	AIT006.12
5	Define unsupervised learning	Unsupervised learning (clustering): The class labels of training data is unknown Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data.	Remember	CO 4	CLO 13	AIT006.13
6	List the evaluating methods of classification	Accuracy • Speed • Robustness • Scalability Interpretability.	Remember	CO 4	CLO 19	AIT006.19
7	What is decision tree?	Decision tree is a flow-chart-like tree structure • Internal node denotes a test on an attribute • Branch represents an outcome of the test Leaf nodes represent class labels or class distribution	Remember	CO 4	CLO15	AIT006.15
8	Describe the different phases of decision tree generation	A directory is a container that is used to contain folders and file. It organizes files and folders into hierarchical manner	Remember	CO 4	CLO 12	AIT006.12
9	Define Pre pruning	Pre pruning: Halt tree construction early. Do not split a node if this would result in the goodness measure falling below a threshold And difficult to choose an appropriate threshold	Remember	CO 4	CLO 19	AIT006.19

10	Define Post pruning	Post pruning: Remove branches from a "fully grown" tree to get a sequence of progressively pruned trees	Remember	CO 4	CLO 13	AIT006.13
11	Define Hierarchical Storage Management	A hierarchical storage system extends the storage hierarchy beyond primary memory and secondary storage to incorporate tertiary storage	Remember	CO 4	CLO 19	AIT006.19
12	List the different attribute selection methods	Attribute Selection Measures • Information Gain • Gain ratio Gini Index	Remember	CO 4	CLO 15	AIT006.15
13	State the ID3 algorithm for constructing decision tree	ID3 uses Information gain and entropy are used to construct decision tree. It employs a top-down, greedy search through the space of possible branches with no backtracking	Remember	CO 4	CLO 13	AIT006.13
14	Define entropy	Entropy (Info(D)) is a measure to find the homogeneity.	Remember	CO 4	CLO 13	AIT006.13
15	Distinguish Lazy vs. eager learning	Lazy learning (e.g., instance-based learning): Simply stores training data (or only minor processing) and waits until it is given a test tuple Eager learning: Given a set of training set, constructs a classification model before receiving new (e.g., test) data to classify Lazy: less time in training but more time in predicting.	Remember	CO 4	CLO 19	AIT006.19
16	Define regression analysis?	Regression analysis is used to study the relationship between two or more variables. Moreover, the regression technique is used to observe changes in the dependent variable with changes in the independent variables. The parameters in the regression equation are obtained by using least square method	Remember	CO 4	CLO11	AIT006.11
17	Define Bayes' Theorem?	In statistics and probability theory, the Bayes' theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.	Remember	CO 4	CLO 15	AIT006.15
18	State gain ratio?	Information gain ratio is a ratio of information gain to the intrinsic information. It was proposed by Ross Quinlan, to reduce a bias towards multi-valued attributes by taking the number and size of branches into account when choosing an attribute	Remember	CO 4	CLO 19	AIT006.19
19	Define Pre Pruning?	A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples	Remember	CO 4	CLO 19	AIT006.19

20	Define the decision tree?	A decision tree is a graphical representation of possible solutions to a decision based on certain conditions	Remember	CO 4	CLO 12	AIT006.12
21	Define the construction of naïve Bayesian classification	These models are generally used to identify the relationship between the input columns and the predicated columns that are available. This algorithm is widely used during the initial stages of the explorations	Remember	CO 4	CLO15	AIT006.15
22	Differentiate supervised learning and unsupervised learning?	Supervised learning: Supervised learning is the learning of the model where with input variable (say, x) and an output variable (say, Y) and an algorithm to map the input to the output. That is, Y = f(X) Unsupervised learning is where only the input data (say, X) is present and no corresponding output variable is there.	Remember	CO 4	CLO 19	AIT006.19
23	Definition of classification	Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data	Remember	CO 4	CLO 15	AIT006.15
24	Define information gain?	Information gain is the amount of information that's gained by knowing the value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it. The largest information gain is equivalent to the smallest entropy	Remember	CO 4	CLO 13	AIT006.13
25	What is index block	Each file has its own index block, which is an array of disk-block addresses. The ith entry in the index block points to the fth block of the file.	Remember	CO 4	CLO 15	AIT006.15
26	Define the if-then rules for classification?	Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following from − IF condition THEN conclusion	Remember	CO 4	CLO 13	AIT006.13
27	State Gini index?	The Gini coefficient measures the inequality among values of a frequency distribution .A Gini coefficient of zero expresses perfect equality where all values are the same (for example, where everyone has an exactly equal income). A Gini coefficient of one (100 on the percentile scale) expresses maximal inequality among values (for example where only one person has all the income)	Remember	CO 4	CLO 13	AIT006.13
28	Define Prediction	Prediction is nothing but finding out the knowledge or some pattern from the large amounts of data	Remember	CO 4	CLO 13	AIT006.13
29	List out The Issues Regarding Classification And Prediction?	Data cleaning , Relevance analysis , Data transformation, Comparing classification methods , Predictive accuracy ,Speed, Robustness, Scalability ,Interpretability	Remember	CO 4	CLO 12	AIT006.12
30	What Is Attribute Selection Measure?	The information Gain measure is used to select the test attribute at each node in the decision tree. Such a	Remember	CO 4	CLO 15	AIT006.15

		measure is referred to as an attribute selection measure or a measure of the goodness of split				
31	What Is The Use Of Regression?	Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula	Remember	CO 4	CLO 15	AIT006.15
UNIT - V						
1	What is a clustering	Clustering is the process of making a group of abstract objects into classes of similar objects. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.	Remember	CO 5	CLO 20	AIT006.21
2	List out the Applications of clustering	Pattern Recognition • Spatial Data Analysis • Create thematic maps in GIS by clustering feature spaces • Detect spatial clusters or for other spatial mining tasks • Image Processing • Economic Science (especially market research) • WWW • Document classification Cluster Weblog data to discover groups of similar access patterns	Remember	CO 5	CLO 20	AIT006.24
3	What is dissimilarity/ Similarity metric?	The similarity between two objects is a numeral measure of the degree to which the two objects are alike. The dissimilarity between two objects is the numerical measure of the degree to which the two objects are different	Remember	CO 5	CLO 20	AIT006.21
4	Define data matrix	A Data Matrix is a two-dimensional barcode consisting of black and white "cells" or modules arranged in either a square or rectangular pattern, also known as a matrix. The information to be encoded can be text or numeric data	Remember	CO 5	CLO 20	AIT006.24
5	Define dissimilarity matrix	The Dissimilarity matrix is a matrix that expresses the similarity pair to pair between two sets. It's square and symmetric. The diagonal members are defined as zero, meaning that zero is the measure of dissimilarity between an element and itself. It is a one mode function	Remember	CO 5	CLO 20	AIT006.24
6	Give different types of data in cluster analysis?	Interval-scaled variables • Binary variables • Nominal, ordinal, and ratio variables Variables of mixed types	Remember	CO 5	CLO 20	AIT006.21
7	State different types of similarity or dissimilarity functions?	Distances are normally used to measure the similarity or dissimilarity between two data objects. Some of the functions are: 1. Minkowski Distancce 2. Manhattan distance Euclidean distance	Remember	CO 5	CLO20	AIT006.21

8	List out the properties of distance function?	Properties d(i,j) >=0
		d(i,i) = 0
		d(i,j) = d(j,i)
		d(i,j) <= d(i,k) + d(k,j)
			Remember	CO 5	CLO 20	AIT006.21

9	List different clustering approaches?	Partitioning approach Hierarchical approach Density-based approach Model-based: Grid-based approach: Frequent pattern-based User-guided or constraint-based	Remember	CO 5	CLO 19	AIT006.21
10	List different methods to calculate the distance between Clusters	Single link Complete link Average Centroid Medoid	Remember	CO 5	CLO 20	AIT006.22
11	List the applications of outlier analysis	Credit card fraud detection Telecom fraud detection Customer segmentation Medical analysis	Remember	CO 5	CLO 20	AIT006.21
12	Differentiate agglomerative and divisive hierarchical clustering	Agglomerative Hierarchical clustering method allows the clusters to be read from bottom to top and it follows this approach so that the program always reads from the sub-component first then moves to the parent whereas divisive uses top-bottom approach in which the parent is visited first then the child	Remember	CO 5	CLO 19	AIT006.23
13	List out the requirements of cluster analysis?	• scalability • dealing with different types of attributes • discovering clusters with arbitrary shape • minimal requirements for domain knowledge to determine input parameters • ability to deal with noise and outliers • insensitivity to order of input records • high dimensionality interpretability and usability	Remember	CO 5	CLO 20	AIT006.21
14	Define Binary variables? And what are the two types of binary variables?	Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights	Remember	CO 5	CLO 20	AIT006.23
15	Define CLARA and CLARANS?	CLARANS(Cluster Large Applications based on Randomized Search) to improve the quality of CLARA we go for CLARANS.	Remember	CO 5	CLO 13	AIT006.13
16	Define Chameleon method	Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster	Remember	CO 5	CLO 12	AIT006.12
17	Define Outlier Detection	Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data	Remember	CO 5	CLO 19	AIT006.19

18	What Is Sequence Clustering Algorithm?	Sequence clustering algorithm collects similar or related paths, sequences of data containing events. The data represents a series of events or transitions between states in a dataset like a series of web clicks. The algorithm will examine all probabilities of transitions and measure the differences, or distances, between all the possible sequences in the data set	Remember	CO 5	CLO 12	AIT006.12
19	What Are Interval Scaled Variables?	Interval scaled variables are continuous measurements of linear scale. For example, height and weight, weather temperature or coordinates for any cluster	Remember	CO 5	CLO15	AIT006.15
20	Define Density Based Method	Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.	Remember	CO 5	CLO 15	AIT006.15
21	What do u mean by partitioning method?	In partitioning method a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster	Remember	CO 5	CLO 15	AIT006.15
22	What Is An Index	Indexes of SQL Server are similar to the indexes in books. They help SQL Server retrieve the data quicker. Indexes are of two types. Clustered indexes and non-clustered indexes	Remember	CO 5	CLO 13	AIT006.13
23	Define visual data mining	Visual data mining discovers implicit and useful knowledge from large data sets using data and/or knowledge visualization techniques	Remember	CO 5	CLO 15	AIT006.15
24	Define multi- media database	Multimedia database is the collection of interrelated multimedia data that includes text, graphics (sketches, drawings), images, animations, video, audio etc and have vast amounts of multisource multimedia data	Remember	CO 5	CLO 13	AIT006.13
25	Define data objects	Data objects can also be referred to as samples, examples, instances, data points, or objects. If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes	Remember	CO 5	CLO11	AIT006.11
26	Difference between time series and sequential data	A time series is a sequence taken at successive equally spaced points in time and it is not the only case of sequential data. In the latter the order is defined by the dimension of time. There are other cases of sequential data as data from text documents, where you can take into account the order of the terms or biological data	Remember	CO 5	CLO 15	AIT006.15
27	List the heuristic method of partitioning clustering	k-means and k-medoids algorithms PAM CLARA CLARANS	Remember	CO 5	CLO 19	AIT006.19

28	Define Hierarchical Clustering?	Hierarchical clustering uses distance matrix as clustering criteria. This method does not require the number of clusters k as an input, but needs a termination condition	Remember	CO 5	CLO 19	AIT006.19
29	List different Hierarchical Clustering algorithms	AGNES-Agglomerative Nesting DIANA-Divisive Analysis	Remember	CO 5	CLO 19	AIT006.19
30	What are outliers	The set of objects are considerably dissimilar from the remainder of the data	Remember	CO 5	CLO 19	AIT006.19

Signature of the Faculty **Signature of the HOD**