MUTHAYAMMAL ENGINEERING COLLEGE
(An Autonomous Institution)
(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu.

MUST KNOW CONCEPTS

MKC

MCA

2021-22

Course Code & Course Name : 19CAC16 & Data Mining and Data Warehousing

Year/Sem/Sec : II / III /-

S.No.	Term	Notation (Symbol)	Concept / Definition / Meaning / Units / Equation / Expression	Units
Unit-I : Data Mining & Data Preprocessing				
1.	Data Mining	-	Extracting or “mining” knowledge from large amounts of data. Data mining is a step in the process of knowledge discovery.	I
2.	Transactional Databases	-	It consists of a file where each record represents a transaction in a given Database.	I
3.	Advantages of Data Mining	-	1. Increasing revenue. 2. Acquiring new customers. 3. Improving cross-selling and up-selling. 4. Detecting fraud.	I
4.	Data Mining Functionalities	-	Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.	I
5.	Descriptive mining	-	Descriptive mining tasks characterize the general properties of the data in the database.	I
6.	Predictive mining	-	Predictive mining tasks perform inference on the current data in order to make predictions.	I
7.	Data Mining Applications	-	1.Financial Data Analysis 2.Retail Industry 3.Telecommunication Industry 4.Biological Data Analysis 5.Other Scientific Applications 6.Intrusion Detection	I
8.	Data Pre-processing	-	Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining.	I
9.	Data Cleaning	-	Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.	I

10.	Data Integration	-	Integration of multiple databases, data cubes, or files.	I
11.	Data Transformation	-	Transform the data in appropriate forms suitable for mining process.	I
12.	Data Reduction	-	Data reduction techniques are applied to obtain a reduced representation of the data set that is much smaller in volume.	I
13.	Data Compression	-	Data compression is the process of encoding, restructuring or otherwise modifying data in order to reduce its size.	I
14.	Challenges in Data Mining	-	1. Security and Social Challenges 2. Noisy and Incomplete Data 3. Distributed Data 4. Complex Data 5. Performance	I
15.	Relational databases	-	A relational databases is a collection of tables, each of which is assigned a unique name.	I
16.	Incomplete Data	-	Lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.	I
17.	Noisy Data	-	Containing Random errors or outliers.	I
18.	Inconsistent Data	-	Containing discrepancies in codes or names.	I
19.	Need for Data Pre - processing	-	Yes Data Pre - Processing is need to check the data quality.	I
20.	Steps in the data mining process	-	1. Data Cleaning 2. Data Integration 3. Data Transformation 4. Data Reduction	I
21.	Intrusion Detection	-	Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources.	I
22.	Pattern evaluation	-	Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interesting measures.	I
23.	knowledge representation	-	Knowledge representation techniques are used to present the mined knowledge to the user.	I
24.	Dimensionality Reduction	-	This reduce the size of data by encoding mechanisms. It can be lossy or lossless.	I
25.	Data Discretization	-	This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.	I
Unit-II : Association Rule Mining And Classification Basics				
26.	Association rule	-	Association rule finds interesting association or correlation relationships among a large set of data items which is	II

			used for decision-making processes.	
27.	Association rules mined from large databases	-	• Find all frequent itemsets. • Generate strong association rules from the frequent itemsets.	II
28.	Frequent pattern mining	-	Frequent pattern mining searches for recurring relationships in a given data set.	II
29.	Strong Association Rules	-	Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_ conf) are called Strong Association Rules.	II
30.	Apriori Algorithm	-	Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent item sets for Boolean association rules	II
31.	Principles of Apriori Algorithm	-	Any subset of a Frequent Item set must be Frequent.	II
32.	Support	-	Support is an association rule interestingness measure.	II
33.	Confidence	-	Confidence is the ratio of the number of transactions.	II
34.	Frequent itemset	-	A set of items is referred to as itemset. An itemset that contains k items is called k-itemset	II
35.	FP Algorithm	-	Frequent - Pattern Growth Algorithm.	II
36.	Apriori Rule	-	An itemset that satisfies minimum support is referred to as frequent itemset.	II
37.	Frequent Pattern Tree	-	Frequent Pattern Tree is a tree-like structure that is made with the initial itemsets of the database.	II
38.	Purpose of the FP tree	-	The purpose of the FP tree is to mine the most frequent pattern. Each node of the FP tree represents an item of the itemset.	II
39.	Advantages of FP Growth Algorithm	-	1.The pairing of items is not done in this algorithm and this makes it faster. 2.The database is stored in a compact version in memory. 3.It is efficient and scalable for mining both long and short frequent patterns.	II
40.	Disadvantages of FP-Growth Algorithm	-	1.FP Tree is more cumbersome and difficult to build than Apriori. 2.It may be expensive. 3.When the database is large, the algorithm may not fit in the shared memory.	II
41.	Constraint-Based Association Mining	-	A data mining process may uncover thousands of rules from a given set of data, most of which end up being unrelated or uninteresting to the users.	II

42.	Multilevel association rule	-	Association rule generated from mining data at multiple level of abstraction.	II
43.	Multidimensional association rule	-	Association rule that involve two or more dimensions or predicates can be refereed.	II
44.	Applications of Association rule mining	-	Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification.	II
45.	Advantages of Dimensional modeling	-	• Predictable, standard framework • Understandable	II
46.	Quantitative association rule	-	Quantitative attribute are numeric and have an implicit ordering among values eg.,age ,income, price.	II
47.	Meta Rules	-	Meta rules allow users to specify the syntactic form of rules that they are interested in mining. It is based on the analyst's experience, expectations.	II
48.	Quantitative association rule	-	Quantitative attribute are numeric and have an implicit ordering among values eg.,age ,income, price	II
49.	Support	-	(A=>B) = P(AUB)	II
50.	Confidence	-	(A=>B) = P(B/A)	II
Unit-III : Classification And Prediction Techniques				
51.	Classification	-	Classification is a data mining technique which predicts categorical class labels classifies data (constructs a model) based on the training set and the values (class labels).	III
52.	Model construction	-	It describing a set of predetermined classes.	III
53.	Model Usage	-	For classifying future or unknown objects.	III
54.	Decision tree	-	1. A flow-chart-like tree structure. 2. Internal node denotes a test on an attribute node (non - leaf node) denotes a test on an attribute.	III
55.	Attribute Selection Measures	-	An attribute selection measure is a heuristic for selecting the splitting criterion.	III
56.	Bayesian Classification	-	Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.	III
57.	Bayes' Theorem	-	$P(H\|X) = \frac{P(X\|H)P(H)}{P(X)}.$	III
58.	Components of Bayesian Belief Networks	-	Two components—a directed acyclic graph and a set of conditional probability tables.	III

59.	2 phases involved in Decision Tree	-	• Tree construction • Tree pruning	III
60.	Rule Based Classification	-	Rule Based Classification is implemented by using IF - THEN Rules.	III
61.	Rules in Rule Based Classification	-	R: IF age = youth AND student = yes THEN buys_computer = yes.	III
62.	Rule Extraction from a Decision Tree	-	1. Rules are easier to understand than large trees. 2.One rule is created for each path from the root to a leaf.	III
63.	Classification by Back Propagation	-	It is a neural network learning algorithm.	III
64.	Neural Network	-	A set of connected input/output units where each connection has a weight associated with it.	III
65.	Support Vector Machines(SVM)	-	A new classification method for both linear and nonlinear data. It uses a nonlinear mapping to transform the original training data into a higher dimension.	III
66.	K-Nearest-Neighbor Classifiers	-	Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.	III
67.	Case-based reasoning (CBR)	-	Case-based reasoning (CBR) classifiers use a database of problem solutions to solve new problems.	III
68.	Internal Node	-	Denotes a test on an attribute	III
69.	Regression Types	-	Linear and multiple regression Non-linear regression	III
70.	Prediction is different from classification	-	Classification refers to predict categorical class label. Prediction models continuous-valued functions.	III
71.	Prediction is similar to classification	-	▪ First, construct a model ▪ Second, use model to predict unknown value	III
72.	Two common approaches to tree pruning	-	1. Pre - Pruning 2. Post - Pruning	III
73.	Tree pruning	-	Identify and remove branches that reflect noise or outliers	III
74.	Uses of Decision trees	-	Used for exploration of data set and business problems.	III
75.	Decision tree pruning	-	Once tree is constructed some modification to the tree might be needed to improve the performance of the tree.	III

Unit-IV : Clustering Techniques				
76.	Cluster	-	A collection of data objects. Similar (or related) to one another within the same group. Dissimilar (or unrelated) to the objects in other groups.	IV
77.	Applications of Clustering	-	1. Biology 2. Information retrieval 3. Marketing 4. Economic Science	IV
78.	What is Good Clustering	-	A good clustering method will produce high quality clusters.	IV
79.	Types of Data in Cluster Analysis	-	1. Interval-Scaled variables 2. Binary variables 3. Nominal, Ordinal, and Ratio variables 4. Variables of mixed types	IV
80.	Binary variables	-	A binary variable is a variable that can take only 2 values.	IV
81.	Partitioning Methods	-	Construct various partitions and then evaluate them by some Condition, e.g., minimizing the sum of square errors.	IV
82.	Hierarchical Methods	-	Create a hierarchical decomposition of the set of data (or objects) using some Condition.	IV
83.	Density-Based Methods	-	Based on connectivity and density functions.	IV
84.	Grid-Based Methods	-	Based on a multiple-level granularity structure.	IV
85.	Model-Based Clustering Methods	-	A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other.	IV
86.	K-means	-	Each cluster is represented by the center of the cluster.	IV
87.	K- medoids	-	k-medoids or PAM (Partition around medoids) : Each cluster is represented by one of the objects in the cluster.	IV
88.	Dendrogram	-	Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.	IV
89.	DIANA	-	DIANA (Divisive Analysis). Introduced in Kaufmann and Rousseeuw (1990). Implemented in statistical analysis packages, e.g., Splus.	IV
90.	STING Clustering Method	-	Each cell at a high level is partitioned into a number of smaller cells in the next lower level.	IV
91.	STING	-	A Statistical Information Grid Approach. The spatial area is divided into rectangular cells.	IV

92.	Wavelet transform	-	A signal processing technique that decomposes a signal into different frequency sub-band.	IV
93.	Dimensionality reduction	-	it is more effective to construct a new space instead of using some sub spaces of the original data.	IV
94.	Constraint-Based Cluster Analysis	-	A constraint refers to the user expectation or the properties of desired clustering results.	IV
95.	Concept hierarchy	-	A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level concepts.	IV
96.	Methods of Outlier Detection	-	1. The statistical approach 2. The distance-based approach 3.The density-based local outlier approach 4. The deviation-based approach	IV
97.	Statistical Approach	-	This approach assumes a distribution for the given data set and then identifies outliers with respect to the model using a discordancy test.	IV
98.	Distance-Based Approach	-	This approach generalizes the ideas behind discordancy testing for various standard.	IV
99.	Deviation-Based Approach	-	This approach identifies outliers by examining the main characteristics of objects in a group.	IV
100.	Outlier	-	Data objects that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.	IV
Unit-V : Data Warehouse				
101.	Data Warehouse	-	A data warehouse is a subject-oriented, time-variant and nonvolatile collection of data in support of management's decision-making process.	V
102.	Data mart	-	A data mart is a segment of a data warehouse that can provide data for reporting and analysis on a section, unit, department or operation in the company.	V
103.	Meta Data	-	It is data about data. It is used for maintaining, managing and using the data warehouse.	V
104.	Inter query Parallelism	-	Different server threads or processes handle multiple requests at the same time.	V
105.	Intra query Parallelism	-	Form of parallelism decomposes the serial SQL query into lower level operations.	V

106.	Star schema	-	Star schema has one large central table (fact table) and a set of smaller tables (dimensions) arranged in a radial pattern around the central table.	V
107.	Snowflake schema	-	Result of decomposing one or more of the dimensions.	V
108.	OLAP	-	OLAP is a powerful technology for data discovery, including capabilities for limitless report viewing.	V
109.	Operational Databases	-	Organizations maintain large database that are updated by daily transactions are called operational databases.	V
110.	Facts	-	Facts are numerical measures. Facts can also be considered as quantities by which can analyze the relationship between dimensions.	V
111.	Data Warehouse tools	-	OLAP(On-Line Analytic Processing) *ROLAP (Relational OLAP) *End User Data Access tool *Ad Hoc Query tool *Data Transformation services *Replication	V
112.	End User Data Access tool	-	End User Data Access tool is a client of the data warehouse. In a relational data warehouse.	V
113.	Characteristics of data warehouse	-	*Available *Integrated *Subject Oriented *Not Dynamic	V
114.	Advantages of a data modeling tool	-	1.Integrates the data warehouse model with other corporate data models. 2.Helps assure consistency in naming.	V
115.	Dimensional Modeling	-	Dimensional modeling is a logical design technique that seeks to present the data in a standard framework.	V
116.	Advantages of Dimensional modeling	-	1.Ease of use. 2.High performance	V
117.	OLTP	-	Online Transaction Processing	V
118.	OLAP	-	Online Analytical Processing	V
119.	OLTP	-	If an on-line operational database systems is used for efficient retrieval, efficient storage and management of large amounts of data, then the system is said to be on-line transaction processing.	V
120.	How a database design is represented in OLTP systems	-	Entity-relation model.	V
121.	How a database design is represented in OLAP systems	-	○ Star schema ○ Snowflake schema ○ Fact constellation schema	V

122.	Data cube	-	It consists of a large set of facts (or) measures and a number of dimensions.	V
123.	Dimensions	-	Dimensions are the entities (or) perspectives with respect to an organization for keeping records and are hierarchical in nature.	V
124.	Dimension Table	-	A dimension table is used for describing the dimension. (e.g.) A dimension table for item may contain the attributes item_ name, brand and type.	V
125.	Fact Table	-	Fact table contains the name of facts (or) measures as well as keys to each of the related dimensional tables.	V
Placement Questions				
126.	Linked cube	-	Linked cube in which a sub-set of the data can be analyzed into detail.	
127.	BUS Schema	-	BUS Schema is composed of a master suite of confirmed dimension and standardized definition if facts.	
128.	Conformed fact	-	Conformed dimensions are the dimensions, which can be used across multiple Data Marts in combination with multiple facts tables accordingly.	
129.	Types of data warehousing	-	1. Enterprise Data warehousing 2. ODS (Operational Data Store) 3. Data Mart	
130.	Fact	-	Fact is key performance indicator to analyze the business.	
131.	Dimension	-	Dimension is used to analyze the fact.	
132.	Measure	-	Without dimension there is no meaning for fact.	
133.	Cluster analysis	-	The process of grouping a data objects into classes of similar objects.	
134.	Outlier Analysis	-	Dissimilar Data objects are outliers.	
135.	Data cleaning	-	Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.	
136.	Roll up down	-	The roll up operation called as drill up operation. It performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.	
137.	Data reduction	-	Obtains reduced representation in volume but produces the same or similar analytical results.	
138.	Support	-	(A=>B) = P(AUB)	
139.	Confidence	-	(A=>B) =P(B/A)	
140.	Applications of Association rule mining	-	Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification.	

141.	Multilevel association rule	-	Multilevel association rule generated from mining data at multiple level of abstraction.	
142.	Partial Backup	-	A Partial Backup is any operating system backup short of a full backup, taken while the database is open or shut down	
143.	Full backup	-	A full backup is an operating system backup of all data files, on- line redo log files and control file that constitute ORACLE database and the parameter.	
144.	Mixed-effect models	-	For analyzing grouped data, i.e. data that can be classified according to one or more grouping variables.	
145.	Intrusion Detection System	-	Defined as the tools, methods, and resources to help identify, assess, and report unauthorized or unapproved network activity.	
146.	Honey pots	-	Servers or systems setup to gather information regarding an attacker of intruder into networks or systems.	
147.	WEKA	-	Waikato Environment for Knowledge Analysis.	
148.	Examples of Data Mining Systems	-	SGI Mine Set DB Miner	
149.	Binary variables	-	A binary variable is a variable that can take only 2 values.	
150.	Tree construction	-	At start, all the training examples are at the root. Partition examples recursively based on selected attributes.	

Faculty Prepared **Signature**

Mr.S.Nithyananth

HoD