



# INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

## INFORMATION TECHNOLOGY

### TUTORIAL QUESTION BANK

<b>Course Title</b>	<b>DATA WAREHOUSING AND DATA MINING</b>			
<b>Course Code</b>	<b>AIT006</b>			
<b>Programme</b>	<b>B.Tech</b>			
<b>Semester</b>	<b>VI</b>			
<b>Course Type</b>	<b>Core</b>			
<b>Regulation</b>	<b>IARE - R16</b>			
<b>Course Structure</b>	Lectures	Tutorials	Practical	Credits
	3	1	3	4
<b>Course Coordinator</b>	<b>Mr.Ch Suresh Kumar Raju, Assistant Professor, Dept. of CSE</b>			
<b>Course Faculty</b>	<b>Mr. A Praveen, Assistant Professor, Dept. of IT</b>			

#### 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.

#### COURSE OUTCOMES (COs):

CO 1	Understand Data Mining concepts and knowledge discovery process
CO 2	Apply task related attribute selection and Data preprocessing techniques
CO 3	Explore on decision tree construction and attribute selection
CO 4	Understand the classification problem and Bayesian classification
CO 5	Explore on different hierarchical based methods, grid based and Model based methods.

#### COURSE LEARNING OUTCOMES:

At the end of the course the students are able to:

S. No	Description
AIT006.01	Learn data warehouse principles and find the differences between relational databases and data warehouse.
AIT006.02	Explore on data warehouse architecture and its components
AIT006.03	Learn Data warehouse schemas
AIT006.04	Distinguish different OLAP Architectures
AIT006.05	Understand Data Mining concepts and knowledge discovery process
AIT006.06	Explore on Data preprocessing techniques
AIT006.07	Apply task related attribute selection and transformation techniques
AIT006.08	Understand the Association rule mining problem

AIT006.09	Illustrate the concept of Apriori algorithm for finding frequent items and generating association rules.
AIT006.10	Explore different representations of frequent item sets.
AIT006.11	Understand the classification problem and decision tree concept
AIT006.12	Understand the classification problem and Bayesian classification
AIT006.13	Illustrate the rule based and back propagation classification algorithms
AIT006.14	Illustrate the rule based and back propagation classification algorithms
AIT006.15	Understand the Clustering Analysis.
AIT006.16	Understand the Types of data and categorization of major clustering methods
AIT006.17	Explore on partition algorithms for clustering.
AIT006.18	Explore on different hierarchical based methods and different density based methods.
AIT006.19	Understand grid based and Model based methods.
AIT006.20	Understand the outlier analysis

S. No	Question	Blooms Taxonomy Level	Course Outcomes	Course learning outcomes
<b>UNIT – I</b>				
<b>DATA WAREHOUSING</b>				
<b>PART – A (Short Answer Questions)</b>				
1	Describe online analytical processing.	Remember	CO 1	AIT006.03
2	List out the key features of data warehouse.	Understand	CO 1	AIT006.03
3	State data mart.	Remember	CO 1	AIT006.03
4	State enterprise warehouse.	Remember	CO 1	AIT006.03
5	State repository.	Remember	CO 1	AIT006.04
6	State metadata.	Remember	CO 1	AIT006.04
7	List out various multidimensional data models.	Understand	CO 1	AIT006.04
8	Describe about the star schema?	Understand	CO 1	AIT006.04
9	Describe the snowflake schema?	Understand	CO 1	AIT006.04
10	Describe about the fact constellation model?	Remember	CO 1	AIT006.05
11	List out the OLAP operations.	Understand	CO 1	AIT006.01
12	Express what is slice and dice operation?	Understand	CO 1	AIT006.01
13	Describe Pivot operation?	Remember	CO 1	AIT006.01
14	Describe concept hierarchy with an example.	Understand	CO 1	AIT006.01
15	State the various views of data warehouse design?	Understand	CO 1	AIT006.01
16	Describe Relational OLAP(ROLAP) server?	Remember	CO 1	AIT006.03
17	Describe about the Multidimensional OLAP(MOLAP) server?	Understand	CO 1	AIT006.03
18	State what is Hybrid OLAP(HOLAP) server?	Understand	CO 1	AIT006.04
19	Describe about the Data warehouse?	Remember	CO 1	AIT006.03
20	List out the uses of concept hierarchy?	Remember	CO 1	AIT006.04
<b>Part - B (Long Answer Questions)</b>				
1	Distinguish between operational database systems and data warehousing?	Understand	CO 1	AIT006.08
2	“Data warehouse is a subject oriented, integrated, time variant and nonvolatile collection of data” illustrate?	Understand	CO 1	AIT006.07
3	Describe the reasons why have a separate data warehouse?	Understand	CO 1	AIT006.01
4	Describe slice and pivot operations on data cube with a neat sketch?	Remember	CO 1	AIT006.01
5	Illustrate the efficient processing of OLAP queries?	Understand	CO 1	AIT006.05
6	Describe the data warehouse applications?	Understand	CO 1	AIT006.04
7	Describe the concept of measures with an example?	Understand	CO 1	AIT006.03
8	Describe various types of OLAP Servers?	Remember		AIT006.01
9	Describe the data warehouse Back-End Tools?	Remember	CO 1	AIT006.04
10	Illustrate the three-tier architecture of a data warehouse with a neat sketch.	Understand	CO 1	AIT006.0 3
11	Describe about Metadata Repository?	Understand	CO 1	AIT006.05
12	Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube.	Understand	CO 1	AIT006.04
13	Explore about the data warehouse implementation with an example?	Understand	CO 1	AIT006.05
14	Design a star schema model by using below data. Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.	Understand	CO 1	AIT006.04
15	Illustrate difference between the three main types of data warehouse usage: information processing, analytical processing, and data mining?	Remember	CO 1	AIT006.03
16	Compare Enterprise warehouse, data mart, and virtual warehouse?	Understand	CO 1	AIT006.04
17	Describe about Data Warehouse Design Process?	Understand	CO 1	AIT006.04

18	Describe about the process of Bit-map indexing OLAP data with an example.	Understand	CO 1	AIT006.04
19	Describe about the process of Join-indexing OLAP data with an example.	Understand	CO 1	AIT006.04
20	Describe about the Efficient Data Cube Computation method?	Understand	CO 1	AIT006.04
<b>Part - C (Critical Thinking Questions)</b>				
1	Design a fact constellation schema model for the following data: Sales are considered along four dimensions:time, item, branch, andlocation. The schema containsa central fact table forsalesthat contains keys to each of the four dimensions, along with two measures:dollarssoldandunitssold Theshippingtable has five dimensions, or keys— itemkey,timekey, shipperkey, fromlocation, andtolocation—and two measures—dollarscost	Understand	CO 1	AIT006.03
2	Suppose that a data warehouse contains 20 dimensions, each with about five levels of granularity. (a) Users are mainly interested in four particular dimensions, each having three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to efficiently support this preference? (b) At times, a user may want to drill through the cube, down to the raw data for one or two particular dimensions. How would you support this feature?	Understand	CO 1	AIT006.03
3	Suppose that a data warehouse for Big University consists of thefollowing four dimensions: student, course, semester, and instructor, andtwoMeasures count and average grade. When at the lowest conceptuallevel (e.g., for a given student, course, semester, and instructor combination), the average grade measure stores the actual course grade of the student. At higher combination. (a)Draw a snowflake schema diagram for the datawarehouse	Understand	CO 1	AIT006.03
4	Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own chargerate. Write the following (a)Draw a star schema diagram for the datawarehouse.	Understand	CO 1	AIT006.03
5	State why, for the integration of multiple heterogeneous information sources, manycompanies in industry prefer theupdate-driven approach(which constructs and usesdata warehouses), rather than thequery-driven approach(which applies wrappers andintegrators). Describe situations where the query-driven approach is preferable to theupdate-driven approach.	Remember	CO 1	AIT006.03
6	Suppose that a data warehouse consists of the four dimensions, date, spectator, location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own chargerate. Write the following (a)Draw a snow-flake schema diagram for the datawarehouse.	Understand	CO 1	AIT006.03
7	In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP), or by a hybrid database technique (HOLAP). (a) Briefly describe each implementation technique.	Remember	CO 1	AIT006.03
8	In data warehouse technology, a multiple dimensional view can be implemented by a relational database technique (ROLAP), or by a multidimensional database technique (MOLAP), or by a hybrid database technique (HOLAP). (b) For each technique, explain how each of the following	Remember	CO 1	AIT006.03

	functions may be implemented: i. The generation of a data warehouse (including aggregation) ii. Roll-up iii. Drill-down Which implementation techniques do you prefer, and why?			
9	Briefly compare the following concepts. You may use an example to Describe your points. (a) Snowflake schema, fact constellation, star schema model (b) Data cleaning, data transformation, refresh	Remember	CO 1	AIT006.03
10	Design a star schema model for the following data. Sales are considered along four dimensions:time, item, branch, andlocation. The schema containsa central fact table forsalesthat contains keys to each of the four dimensions, along with two measures:dollarssoldandunitsold. To minimize the size of the fact table,dimension identifiers (e.g.,timekeyanditemkey) are system-generated identifiers.	Remember	CO 1	AIT006.03

**UNIT – II  
DATA MINING**

**Part – A (Short Answer Questions)**

1	Enumerate about data mining.	Remember	CO 2	AIT006.05
2	List the steps involved in knowledge discovery in data (or) KDD method?	Understand	CO 2	AIT006.06
3	Distinguish between data mining and data warehouse.	Understand	CO 2	AIT006.05
4	Express any three functionality of data mining.	Remember	CO 2	AIT006.06
5	List out major issues in data mining	Understand	CO 2	AIT006.07
6	Describe about the spatial temporal databases?	Remember	CO 2	AIT006.05
7	Enumerate about relational databases?	Understand	CO 2	AIT006.06
8	State object –oriented Databases?	Understand	CO 2	AIT006.06
9	Describe about the spatial databases?	Understand	CO 2	AIT006.05
10	Contrast heterogeneous databases and legacy databases?	Understand	CO 2	AIT006.05
11	Distinguish classification and Prediction?	Understand	CO 2	AIT006.05
12	Describe transactional data bases?	Remember	CO 2	AIT006.05
13	List out the types of data that can be mined?	Remember	CO 2	AIT006.06
14	Illustrate data objects and attribute types.	Remember	CO 2	AIT006.06
15	Elucidate multidimensional data mining?	Remember	CO 2	AIT006.05
16	Narrate about data characterization?	Remember	CO 2	AIT006.06
17	List outthe reasons for data preprocessing	Understand	CO 2	AIT006.07
18	Demonstrate about the outlier analysis?	Understand	CO 2	AIT006.06
19	List out the steps involved in data preprocessing?	Understand	CO 2	AIT006.07
20	Illustrate about the dimensionality reduction?	Understand	CO 2	AIT006.05

**Part - B (Long Answer Questions)**

1	Describe data mining? In your answer, address the following: (a)Is it hype? (b) Is it a simple transformation of Technology developed from Databases, statistics, and machine learning? (c) Describe how the evolutions of database technology lead to datamining? (d)Describe the steps involved in data mining when viewed as a processof Remember discovery.	Understand	CO 2	AIT006.05
2	Present an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?	Remember	CO 2	AIT006.06
3	Compare between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?	Understand	CO 2	AIT006.07
4	Describe three challenges to data mining regarding data mining Methodology and user interaction issues?	Remember	CO 2	AIT006.05
5	Distinguish between the data warehouses and data mining?	Remember	CO 2	AIT006.05
6	Narrate about the data smoothing techniques?	Understand	CO 2	AIT006.06

7	Illustrate Data Integration and Transformation?	Understand	CO 2	AIT006.06
8	Describe the various data reduction techniques?	Understand	CO 2	AIT006.07
9	Express about data cleaning? Express the different techniques for handling Missing values?	Remember	CO 2	AIT006.05
10	Distinguish between descriptive and predictive data mining?	Understand	CO 2	AIT006.06
11	Express data mining as a step in the process of Knowledge discovery?	Understand	CO 2	AIT006.06
12	Describe briefly Discretization and concept hierarchy generation for numerical data?	Remember	CO 2	AIT006.07
13	Enumerate about the concept hierarchy generation for categorical data?	Understand	CO 2	AIT006.07
14	List out and describe the five primitives for specifying a data mining task?	Understand	CO 2	AIT006.05
15	Demonstrate the issues to considered during data integration?	Understand	CO 2	AIT006.06
16	Describe the following advanced database systems and applications: object- relational databases, spatial databases, text databases, multimedia databases, stream data, the World WideWeb.	Remember	CO 2	AIT006.05
17	Describe why concept hierarchies are useful in data mining.	Remember	CO 2	AIT006.05
18	Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi tight coupling, and tight Coupling. State which approach you think is the most popular, and why?	Remember	CO 2	AIT006.06
19	Illustrate about the Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.	Understand	CO 2	AIT006.06
20	Understand the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 min-max normalization by setting min = 0 and max =1 z-score normalization	Understand	CO 2	AIT006.07
<b>Part – C (Problem Solving and Critical Thinking)</b>				
1	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20,20, 21, 22, 22, 25, 25,25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52,70. Solve the following: (a) Mean of the data?Median? (b) mode of the data? Comment on the data's modality (i.e.bimodal, trimodal etc.). (c) midrange of thedata?	Understand	CO 2	AIT006.05
2	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19,20,20,21,22,22,25,2525, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. Solve the following: (a) Can you find (roughly) the first quartile (Q1) and the thirdquartile (Q3) of thedata? (b) Give the five-number summary of thedata. (c) Show a box plot of thedata. (d) How is a quantile-quantile plot different from a quantileplot?	Understand	CO 2	AIT006.06
3	Use the data for age given above answer the following. (a) Use smoothing by bin means to smooth the above data, using a bindepth of3. Illustrate your steps. Comment on the effect of this technique for the given data (b) How might you determine outliers in thedata? (c) What other methods are there for datasmoothering?	Understand	CO 2	AIT006.05

4	Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result age 23 23 27 27 39 41 47 49 %fat 9.5 26.57.8 17.8 31.4 25.9 27.4 27.2 31.2 age 52 54 54 56 57 58 58 60 %fat 34.6 42.528.8 33.4 30.2 34.1 32.9 41.2 35.7 Examine the following (a) the mean, median and standard deviation of age and%fat. (b) Draw the box plots for age and%fat. (c) Draw a scatter plot and a q-q plot based on these twovariables.	Remember	CO 2	AIT006.06
5	Describe the deference between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi tight coupling, and tight coupling. State which approach you think is the most popular, and why.	Understand	CO 2	AIT006.07
6	Suppose your task as a software engineer at Big University is to design a data mining system to examine the university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and the cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?	Understand	CO 2	AIT006.08
7	Outliers are often discarded as noise. However, one person's garbage could be another's treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Taking fraudulencedetectionas an example, List out the two methods that can be used to detect outliers and Illustrate which one is more reliable.	Understand	CO 2	AIT006.06
9	Examine the following consider the following data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36 , 40, 45, 46, 52, 70. (a) Use min-max normalization to transform the value 35 for age on tothe range [0.0,1.0]. (b) Use z-score normalization to transform the value 35 for age, wherethe standard deviation of age is 12.94years. (c) Use normalization by decimal scaling to transform the value 35 forage. (d) Comment on which method you would prefer to use for the givendata, giving reasons as towwhy.	Remember	CO 2	AIT006.07

**UNIT – III  
ASSOCIATION MINING**

**MID-1**

**Part – A (Short Answer Questions);**

S. No	Question	Blooms Taxonomy Level	Course Outcomes	Course learning outcomes
1.	State association rule?	Remember	CO 3	AIT006.08
2.	State item set?	Remember	CO 3	AIT006.08
3.	State frequent item sets?	Understand	CO 3	AIT006.08
4.	List out the measures of association rules?	Understand	CO 3	AIT006.08
5.	List out the types of association rules?	Understand	CO 3	AIT006.09
6.	List out the principle of APRIORI algorithm?	Understand	CO 3	AIT006.09
7.	State the problem definition for association rules?	Remember	CO 3	AIT006.10
8.	Describe support and minimum support?	Understand	CO 3	AIT006.09



9.	State confidence and minimum confidence for strong association rule?	Understand	CO 3	AIT006.10																				
10.	List out the steps in association rule mining?	Remember	CO 3	AIT006.10																				
11.	Describe the two kinds of closure checking?	Understand	CO 3	AIT006.09																				
12.	Describe the five categories of pattern mining constraints?	Understand	CO 3	AIT006.08																				
13.	List out the techniques of efficiency of Apriori algorithm?	Remember	CO 3	AIT006.09																				
14.	List out the drawbacks of Apriori technique?	Understand	CO 3	AIT006.08																				
<b>Part – B (Long Answer Questions)</b>																								
1	State the terms frequent item sets, closed item sets and association rules?	Remember	CO 3	AIT006.09																				
2	Illustrate which algorithm is an influential algorithm for mining frequent Item sets for boolean association rules? Describe with an example?	Understand	CO 3	AIT006.08																				
3	Describe the different techniques to improve the efficiency of Apriori?	Remember	CO 3	AIT006.09																				
4	Illustrate the FP-growth algorithm with an example?	Understand	CO 3	AIT006.08																				
5	Describe how to mine the frequent item sets using vertical data format?	Understand	CO 3	AIT006.08																				
6	Illustrate about mining multilevel association rules from transaction databases in detail?	Understand	CO 3	AIT006.09																				
7	Describe how to mine the multidimensional association rules from relational databases and data warehouses?	Understand	CO 3	AIT006.09																				
8	Describe briefly about the different correlation measures in association analysis?	Understand	CO 3	AIT006.08																				
9	Illustrate about constraint-based association mining?	Understand	CO 3	AIT006.08																				
10	Describe the Apriori algorithm with example?	Understand	CO 3	AIT006.09																				
<b>Part – C (Problem Solving and Critical Thinking Questions)</b>																								
1	<p>The following</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>T ID</th> <th>List of item ID's</th> </tr> </thead> <tbody> <tr> <td>T100</td> <td>I1, I2, I5</td> </tr> <tr> <td>T200</td> <td>I2, I4</td> </tr> <tr> <td>T300</td> <td>I2, I3</td> </tr> <tr> <td>T400</td> <td>I1, I2, I4</td> </tr> <tr> <td>T500</td> <td>I1, I3</td> </tr> <tr> <td>T600</td> <td>I2, I3</td> </tr> <tr> <td>T700</td> <td>I1, I3</td> </tr> <tr> <td>T800</td> <td>I1, I2, I3, I5</td> </tr> <tr> <td>T900</td> <td>I1, I2, I3</td> </tr> </tbody> </table> <p>(a) List out all frequent item sets using Apriori with Min support count 2</p> <p>(b) List out all of the strong association rules</p>	T ID	List of item ID's	T100	I1, I2, I5	T200	I2, I4	T300	I2, I3	T400	I1, I2, I4	T500	I1, I3	T600	I2, I3	T700	I1, I3	T800	I1, I2, I3, I5	T900	I1, I2, I3	Understand	CO 3	AIT006.08
T ID	List of item ID's																							
T100	I1, I2, I5																							
T200	I2, I4																							
T300	I2, I3																							
T400	I1, I2, I4																							
T500	I1, I3																							
T600	I2, I3																							
T700	I1, I3																							
T800	I1, I2, I3, I5																							
T900	I1, I2, I3																							
3	Illustrate about frequent item set? Write the Apriori algorithm for frequent item set generation? Describe with an example	Understand	CO 3	AIT006.09																				
4	Illustrate about Market basket analysis with suitable example	Understand	CO 3	AIT006.09																				
5	<p>How can we mine multilevel Association rules efficiently using concept hierarchies? Illustrate with an A-priori algorithm for the given dataset below.</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>TID</th> <th>List of items</th> </tr> </thead> <tbody> <tr> <td>T001</td> <td>Milk, dal, sugar, bread</td> </tr> <tr> <td>T002</td> <td>Dal, sugar, wheat, jam</td> </tr> <tr> <td>T003</td> <td>Milk, bread, curd, paneer</td> </tr> <tr> <td>T004</td> <td>Wheat, paneer, dal, sugar</td> </tr> <tr> <td>T005</td> <td>Milk, paneer, bread</td> </tr> <tr> <td>T006</td> <td>Wheat, dal, paneer, bread</td> </tr> </tbody> </table>	TID	List of items	T001	Milk, dal, sugar, bread	T002	Dal, sugar, wheat, jam	T003	Milk, bread, curd, paneer	T004	Wheat, paneer, dal, sugar	T005	Milk, paneer, bread	T006	Wheat, dal, paneer, bread	Understand	CO 3	AIT006.09						
TID	List of items																							
T001	Milk, dal, sugar, bread																							
T002	Dal, sugar, wheat, jam																							
T003	Milk, bread, curd, paneer																							
T004	Wheat, paneer, dal, sugar																							
T005	Milk, paneer, bread																							
T006	Wheat, dal, paneer, bread																							



**MID 11**

<b>Part - A (Short Answer Questions);</b>																								
1	Describe examples for frequent item sets?	Understand	CO 3	AIT006.09																				
2	List out the pruning strategies in mining closed frequent item sets?	Understand	CO 3	AIT006.09																				
3	Describe the join step?	Understand	CO 3	AIT006.08																				
4	Describe the prune step?	Remember	CO 3	AIT006.08																				
5	State how can we mine closed frequent item sets?	Understand	CO 3	AIT006.09																				
6	List out the pruning strategies of closed frequent item sets?	Remember	CO 3	AIT006.09																				
7	Describe the rule of support for item sets?	Understand	CO 3	AIT006.08																				
8	Describe the two kinds of closure checking?	Understand	CO 3	AIT006.08																				
9	List out the techniques to improve the efficiency of Apriori algorithm?	Understand	CO 3	AIT006.09																				
10	Describe the procedure to find association rule from given frequent item sets?	Understand	CO 3	AIT006.08																				
<b>Part – B (Long Answer Questions)</b>			CO 3																					
1	Illustrate the generating association rules from frequent item sets.	Understand	CO 3	AIT006.09																				
2	Illustrate about mining multilevel association rules from transaction databases in detail?	Understand	CO 3	AIT006.08																				
3	Describe multidimensional association rules using static Discretization?	Remember	CO 3	AIT006.08																				
4	Describe what are additional rule constraints to guide mining?	Understand	CO 3	AIT006.09																				
5	Describe, how can we tell which strong association rules are really interesting? Describe with an example?	Understand	CO 3	AIT006.09																				
6	Describe about the correlation analysis using Chi-square?	Remember	CO 3	AIT006.09																				
7	Describe about the Mining closed Frequent Item set	Remember	CO 3	AIT006.08																				
8	Prove that items in a strong association rule may actually be negatively correlated.	Understand	CO 3	AIT006.09																				
9	Describe Association rule mining often generates a large number of rules. Illustrate effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.	Understand	CO 3	AIT006.09																				
<b>Part – C (Problem Solving and Critical Thinking Questions)</b>																								
1	Describe support and confidence by using with following transactional data and find frequent item sets using FP growth tree.  <table border="1" style="margin-left: 40px;"> <thead> <tr> <th>T ID</th> <th>List of item ID's</th> </tr> </thead> <tbody> <tr> <td>T100</td> <td>I1, I2, I5</td> </tr> <tr> <td>T200</td> <td>I2, I4</td> </tr> <tr> <td>T300</td> <td>I2, I3</td> </tr> <tr> <td>T400</td> <td>I1, I2, I4</td> </tr> <tr> <td>T500</td> <td>I1, I3</td> </tr> <tr> <td>T600</td> <td>I2, I3</td> </tr> <tr> <td>T700</td> <td>I1, I3</td> </tr> <tr> <td>T800</td> <td>I1, I2, I3, I5</td> </tr> <tr> <td>T900</td> <td>I1, I2, I3</td> </tr> </tbody> </table>	T ID	List of item ID's	T100	I1, I2, I5	T200	I2, I4	T300	I2, I3	T400	I1, I2, I4	T500	I1, I3	T600	I2, I3	T700	I1, I3	T800	I1, I2, I3, I5	T900	I1, I2, I3	Understand	CO 3	AIT006.09
T ID	List of item ID's																							
T100	I1, I2, I5																							
T200	I2, I4																							
T300	I2, I3																							
T400	I1, I2, I4																							
T500	I1, I3																							
T600	I2, I3																							
T700	I1, I3																							
T800	I1, I2, I3, I5																							
T900	I1, I2, I3																							

2	<p>A database has five transactions. Let min sup = 60% and min conf = 80%.</p> <table border="1" data-bbox="375 296 911 615"> <thead> <tr> <th>TID</th> <th>items bought</th> </tr> </thead> <tbody> <tr> <td>T100</td> <td>{M, OO, N, K, E,Y}</td> </tr> <tr> <td>T200</td> <td>{D, O, N, K, E, Y}</td> </tr> <tr> <td>T300</td> <td>{M, A, K,E}</td> </tr> <tr> <td>T400</td> <td>{M, U, C, K,Y}</td> </tr> <tr> <td>T500</td> <td>{C, O, O, K, I,E}</td> </tr> </tbody> </table> <p>Examine the following</p> <p>(d) Find all frequent item sets usingFP-growth.</p> <p>(e) List out all of the strong association rules (with support s and confidence c) matching the following meta rule, where X is a variable representing customers, and item i denotes variables representing items (e.g., “A”, “B”, etc.):</p>	TID	items bought	T100	{M, OO, N, K, E,Y}	T200	{D, O, N, K, E, Y}	T300	{M, A, K,E}	T400	{M, U, C, K,Y}	T500	{C, O, O, K, I,E}	Remember	CO 3	AIT006.08		
TID	items bought																	
T100	{M, OO, N, K, E,Y}																	
T200	{D, O, N, K, E, Y}																	
T300	{M, A, K,E}																	
T400	{M, U, C, K,Y}																	
T500	{C, O, O, K, I,E}																	
3	<p>Compare Apriori and FP growth algorithms for frequent item set mining in transactional databases. Apply these algorithms to the following data:</p> <p>TID LIST OF ITEMS</p> <p>1 Bread, Milk, Sugar, TeaPowder, Cheese, Tomato</p> <p>2 Onion, Tomato, Chillies, Sugar, Milk</p> <p>3 Milk, Cake, Biscuits, Cheese, Onion</p> <p>4 Chillies, Potato, Milk, Cake, Sugar, Bread</p> <p>5 Bread, Jam, Mik, Butter, Chilles</p> <p>6 Butter, Cheese, Paneer, Curd, Milk, Biscuits</p> <p>7 Onion, Paneer, Chilies, Garlic, Milk</p> <p>8 Bread, Jam, Cake, Biscuits, Tomato</p>	Remember	CO 3	AIT006.08														
4	<p>A database has six transactions. Let min-sup = 50% and min-conf =75%.</p> <p>Find all frequent item sets using Apriori algorithm. List all the strong association rules</p> <table border="1" data-bbox="305 1377 956 1577"> <thead> <tr> <th>TID</th> <th>List of items</th> </tr> </thead> <tbody> <tr> <td>T001</td> <td>Pencil, sharpener, eraser, color papers</td> </tr> <tr> <td>T002</td> <td>Color papers, charts, glue sticks</td> </tr> <tr> <td>T003</td> <td>Pencil, glue stick, eraser, pen</td> </tr> <tr> <td>T004</td> <td>Oil pastels, poster colours, correction tape</td> </tr> <tr> <td>T005</td> <td>Whitener, pen , pencil, charts, glue stick</td> </tr> <tr> <td>T006</td> <td>Colour pencils, crayons, eraser, pen</td> </tr> </tbody> </table>	TID	List of items	T001	Pencil, sharpener, eraser, color papers	T002	Color papers, charts, glue sticks	T003	Pencil, glue stick, eraser, pen	T004	Oil pastels, poster colours, correction tape	T005	Whitener, pen , pencil, charts, glue stick	T006	Colour pencils, crayons, eraser, pen	Remember	CO 3	AIT006.08
TID	List of items																	
T001	Pencil, sharpener, eraser, color papers																	
T002	Color papers, charts, glue sticks																	
T003	Pencil, glue stick, eraser, pen																	
T004	Oil pastels, poster colours, correction tape																	
T005	Whitener, pen , pencil, charts, glue stick																	
T006	Colour pencils, crayons, eraser, pen																	
5	<p>Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, Describe with following example</p> <table border="1" data-bbox="305 1692 794 1890"> <thead> <tr> <th>TID</th> <th>List of items</th> </tr> </thead> <tbody> <tr> <td>T001</td> <td>Milk, dal, sugar, bread</td> </tr> <tr> <td>T002</td> <td>Dal, sugar, wheat, jam</td> </tr> <tr> <td>T003</td> <td>Milk, bread, curd, paneer</td> </tr> <tr> <td>T004</td> <td>Wheat, paneer, dal, sugar</td> </tr> <tr> <td>T005</td> <td>Milk, paneer, bread</td> </tr> <tr> <td>T006</td> <td>Wheat, dal, paneer, bread</td> </tr> </tbody> </table>	TID	List of items	T001	Milk, dal, sugar, bread	T002	Dal, sugar, wheat, jam	T003	Milk, bread, curd, paneer	T004	Wheat, paneer, dal, sugar	T005	Milk, paneer, bread	T006	Wheat, dal, paneer, bread	Understand	CO 3	AIT006.08
TID	List of items																	
T001	Milk, dal, sugar, bread																	
T002	Dal, sugar, wheat, jam																	
T003	Milk, bread, curd, paneer																	
T004	Wheat, paneer, dal, sugar																	
T005	Milk, paneer, bread																	
T006	Wheat, dal, paneer, bread																	

**UNIT-IV  
CLASSIFICATION AND PREDICTION**

**Part – A (Short Answer Questions)**

1	State classification?	Understand	CO 4	AIT006.14
---	-----------------------	------------	------	-----------

2	State regression analysis?	Remember	CO 4	AIT006.14
3	List out the steps in data classification?	Understand	CO 4	AIT006.13
4	State training tuple?	Remember	CO 4	AIT006.14
6	Describe accuracy of a classifier?	Remember	CO 4	AIT006.15
7	Distinguish supervised learning and unsupervised learning?	Understand	CO 4	AIT006.14
8	State the decision tree?	Understand	CO 4	AIT006.14
9	State information gain?	Remember	CO 4	AIT006.13
10	State gain ratio?	Understand	CO 4	AIT006.14
11	State Gini index?	Understand	CO 4	AIT006.14
12	Describe tree pruning?	Understand	CO 4	AIT006.14
14	State the construction of naïve Bayesian classification?	Understand	CO 4	AIT006.14
15	Describe the IF-THEN rules for classification?	Understand	CO 4	AIT006.13
16	Describe Decision Tree Induction?	Understand	CO 4	AIT006.12
17	List out the Attribute Selection Measures?	Remember	CO 4	AIT006.14
18	State Bayes' Theorem?	Understand	CO 4	AIT006.13
19	State Naïve Bayesian Classification?	Remember	CO 4	AIT006.12
20	Describe K-Nearest-Neighbor Classifiers?	Understand	CO 4	AIT006.14
<b>Part – B (Long Answer Questions)</b>				
1	Describe about the classification and prediction? Example with an Example?	Understand	CO 4	AIT006.12
2	Illustrate about basic decision tree induction algorithm?	Understand	CO 4	AIT006.14
3	Describe briefly various measures associated with attribute selection?	Understand	CO 4	AIT006.13
4	Summarize how does tree pruning work? What are some enhancements to basic decision tree induction?	Understand	CO 4	AIT006.14
5	Describe how scalable is decision tree induction? Describe?	Understand	CO 4	AIT006.13
6	Describe the working procedures of simple Bayesian classifier?	Remember	CO 4	AIT006.14
7	Describe Bayesian Belief Networks?	Understand	CO 4	AIT006.12
8	Illustrate about k-nearest neighbor classifier and case-based reasoning?	Understand	CO 4	AIT006.13
9	Describe about classifier accuracy? Describe the process of measuring the accuracy of a classifier?	Understand	CO 4	AIT006.12
10	Describe any ideas can be applied to any association rule mining be applied to classification?	Remember	CO 4	AIT006.13
11	Describe about the major issues regarding classifications and predictions?	Understand	CO 4	AIT006.15
12	Distinguish classification and prediction methods?	Understand	CO 4	AIT006.15
13	Describe briefly various measures associated with attribute selection?	Understand	CO 4	AIT006.15
14	Describe training of Bayesian belief networks?	Understand	CO 4	AIT006.15
15	Describe how tree pruning useful in decision tree induction? What isa drawback of using a separate set of tuplesto evaluate pruning?	Understand	CO 4	AIT006.12
16	Describe for a given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the Decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?	Understand	CO 4	AIT006.14
17	Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k- nearest neighbor, case- based reasoning).	Understand	CO 4	AIT006.14
18	Describe an algorithm for k-nearest-neighbor classification given k and n, the number of attributes describing each tuple.	Understand	CO 4	AIT006.13
19	Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii)input parameters that must be specified;and (iii) limitations. (a)k-means (b)k-medoids	Remember	CO 4	AIT006.13

<b>Part – C (Problem Solving and Critical Thinking Questions)</b>				
1	Illustrate why is tree pruning useful in decision tree induction? Describe the drawback of using a separate set of tuples to evaluate pruning?	Understand	CO 4	AIT006.14
2	Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. Describe advantage does(a) have over (b)?	Understand	CO 4	AIT006.14
3	Outline the major ideas of naive Bayesian classification. Describe why is naïve Bayesian classification called “naive”?	Understand	CO 4	AIT006.14
4	Design an efficient method that performs effective naive Bayesian classification over an infinite data stream (i.e., you can scan the data stream only once). If we wanted to discover the evolution of such classification schemes (e.g., comparing the classification scheme at this moment with earlier schemes, such as one from a week ago), Construct modified design would you suggest?	Understand	CO 4	AIT006.14
5	Illustrate K- Nearest neighbor classification-Algorithm and Characteristics with example.	Understand	CO 4	AIT006.14
6	Describe in detail How does the Naïve Bayesian classification works?	Understand	CO 4	AIT006.13
8	State is associative classification? Why is associative classification able to achieve higher classification accuracy than a classical decision tree method? Describe how associative classification can be used for text document classification.	Understand	CO 4	AIT006.12
9	It is difficult to assess classification accuracy when individual data objects may belong to more than one class at a time. In such cases, Describe on what criteria you would use to compare different classifiers modeled after the same data.	Understand	CO 4	AIT006.12

**UNIT-V  
CLUSTERING**

<b>Part - A (Short Answer Questions)</b>				
1	State Clustering?	Remember	CO5	AIT006.18
2	Illustrate the meaning of cluster analysis?	Understand	CO5	AIT006.18
3	Describe the fields in which clustering techniques are used?	Understand	CO5	AIT006.17
4	List out the requirements of cluster analysis?	Remember	CO5	AIT006.18
5	Express the different types of data used for cluster analysis?	Understand	CO5	AIT006.17
6	State interval scaled variables?	Remember	CO5	AIT006.17
7	State Binary variables? And what are the two types of binary variables?	Remember	CO5	AIT006.17
8	State nominal, ordinal and ratio scaled variables?	Remember	CO5	AIT006.17
9	Illustrate mean by partitioning method?	Understand	CO5	AIT006.16
10	State CLARA and CLARANS?	Remember	CO5	AIT006.16
11	State hierarchical method?	Remember	CO5	AIT006.16
12	Distinguish agglomerative and divisive hierarchical clustering?	Understand	CO5	AIT006.06
13	State K-Means method?	Remember	CO5	AIT006.18
14	State Outlier Detection?	Remember	CO5	AIT006.17
20	State Chameleon method?	Remember	CO5	AIT006.19
<b>Part - B (Long Answer Questions)</b>				
1	Illustrate the various types of data in cluster analysis?	Understand	CO5	AIT006.016
2	Describe the categories of major clustering methods?	Understand	CO5	AIT006.17
3	Describe algorithms for k-means and k-medoids?	Understand	CO5	AIT006.16
4	Describe the different types of hierarchical methods?	Understand	CO5	AIT006.17
5	Demonstrate about the following hierarchical methods a) BIRCH b) Chameleon	Understand	CO5	AIT006.17
6	Describe about semi-supervised cluster analysis?	Understand	CO5	AIT006.17
7	Describe about the outlier analysis?	Understand	CO5	AIT006.16

8	State the distance-based outlier? Illustrate the efficient algorithms for mining distance-based algorithm?	Remember	CO5	AIT006.17
9	Describe about the Statistical-based outlier detection?	Understand	CO5	AIT006.15
10	Describe about the distance-based outlier detection?	Remember	CO5	AIT006.15
11	Illustrate about the density-based outlier detection?	Understand	CO5	AIT006.15
12	Demonstrate about the deviation-based outlier detection techniques?	Understand	CO5	AIT006.15
13	Demonstrate about the BIRCH hierarchical methods?	Understand	CO5	AIT006.15
14	standardize the variable by the following: (a) Calculate the mean absolute deviation of age. (b) Calculate the z-score for the first four measurements.		CO5	
15	Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (such as AGNES).	Understand	CO5	AIT006.16
16	Describe why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distance-based outlier detection, density-based local outlier detection, and deviation-based outlier detection.	Understand	CO5	AIT006.16
17	Describe briefly mining of multimedia databases and time series databases.	Understand	CO5	AIT006.16
<b>Part – C (Problem Solving and Critical Thinking Questions)</b>				
1	Given the following measurements for the variable age: 48, 12, 25, 42, 28, 43, 33, 35, 56, 28, standardize the variable by the following: Calculate (a) The mean absolute deviation of age. (b) The z-score for the first four measurements.	Understand	CO5	AIT006.15
2	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): Calculate (a) The Euclidean distance between the two objects. (b) The Manhattan distance between the two objects. (c) The Minkowski distance between the two objects, using $p = 3$ .	Understand	CO5	AIT006.15
3	Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters. A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-means algorithm to show only The three cluster centers after the first round of execution and The final three clusters	Understand	CO5	AIT006.15
4	Describe why is it that BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not? Can you propose some Modifications to BIRCH to help it find clusters of arbitrary shape?	Understand	CO5	AIT006.15
5	Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations. k-means (b) k-medoids (c) CLARA	Understand	CO5	AIT006.17
6	Give a brief note on PAM Algorithm with example and Write the key issue in hierarchical clustering algorithm.	Understand	CO5	AIT006.17
7	List out the different clustering methods? Describe in detail.	Understand	CO5	AIT006.17

8	<p>State K-means algorithm. Apply k-means algorithm with two iterations to form two clusters by taking the initial cluster centers as subjects 1 and 4.</p> <table border="1" data-bbox="412 296 797 516"> <thead> <tr> <th>Subject</th> <th>A</th> <th>B</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>1.0</td> <td>1.0</td> </tr> <tr> <td>2</td> <td>1.5</td> <td>2.0</td> </tr> <tr> <td>3</td> <td>3.0</td> <td>4.0</td> </tr> <tr> <td>4</td> <td>5.0</td> <td>7.0</td> </tr> <tr> <td>5</td> <td>3.5</td> <td>5.0</td> </tr> <tr> <td>6</td> <td>4.5</td> <td>5.0</td> </tr> <tr> <td>7</td> <td>3.5</td> <td>4.5</td> </tr> </tbody> </table>	Subject	A	B	1	1.0	1.0	2	1.5	2.0	3	3.0	4.0	4	5.0	7.0	5	3.5	5.0	6	4.5	5.0	7	3.5	4.5	Understand	CO5	AIT006.19
Subject	A	B																										
1	1.0	1.0																										
2	1.5	2.0																										
3	3.0	4.0																										
4	5.0	7.0																										
5	3.5	5.0																										
6	4.5	5.0																										
7	3.5	4.5																										
9	<p>Given the following measurements for the variable age: 29, 31, 25, 41, 27, 43, 33, 35, 56, 28, standardize the variable by the following:            Calculate            The mean absolute deviation of age.            The z-score for the first three measurements.</p>	Understand	CO5	AIT006.16																								
10	<p>Given two objects represented by the tuples (21, 2, 41, 11) and (21, 1, 32, 6):            Calculate            (a) The Euclidean distance between the two objects. (b) The Manhattan distance between the two objects.            (c) The Minkowski distance between the two objects, using <math>p = 2</math>.</p>	Understand	CO5	AIT006.18																								

**Prepared by:**

Mr.Ch Suresh Kumar Raju , Assistant Professor

**HOD, IT**