# INSTITUTE OF AERONAUTICAL ENGINEERING

**(Autonomous)**

Dundigal, Hyderabad - 500 043

## COMPUTER SCIENCE AND ENGINEERING

### TUTORIAL QUESTION BANK

| | | | | |
|---|---|---|---|---|
| **Course Title** | **DATA WAREHOUSING AND DATA MINING** | | | |
| **Course Code** | **AIT006** | | | |
| **Programme** | B.Tech | | | |
| **Semester** | VI | | | |
| **Course Type** | Core | | | |
| **Regulation** | IARE - R16 | | | |
| **Course Structure** | Lectures | Tutorials | Practical | Credits |
| | 3 | 1 | 3 | 4 |
| **Course Coordinator** | **Mr.Ch Suresh Kumar Raju, Assistant Professor, Dept. of CSE** | | | |
| **Course Faculty** | **Dr. M MadhuBala, Professor, Dept. of CSE**<br>**Dr. D Kishore Babu, Professor, Dept. of CSE**<br>**Mr.Ch Suresh Kumar Raju, Assistant Professor, Dept. of CSE**<br>**Ms. S Swarajyalaxmi, Assistant Professor, Dept. of CSE**<br>**Ms. M GeethaYadav, Assistant Professor, Dept. of CSE** | | | |

### 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 principlesand 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 andcategorization 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  andModel based methods. |
| AIT006.20 | Understand the outlier analysis |

| S. No | Question | Blooms Taxonomy Level | Course Outcomes | Course learning outcomes |
|---|---|---|---|---|
| UNIT – I<br>DATA WAREHOUSING | | | | |
| | PART – A (Short Answer Questions) | | | |
| 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 |
| | Part - B (Long Answer Questions) | | | |
| 1 | Distinguishbetween 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 functionsused 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.<br>Suppose that a data warehouse consists of the three dimensions time, doctor, andpatient, and the two measurescountandcharge, wherechargeis the fee that a doctor charges apatient for a visit. | Understand | CO 1 | AIT006.04 |
| 15 | Illustrate difference between the three main types of data warehouse usage:infor-mation processing,analytical processing, anddata 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 |
| | **Part - C (Critical Thinking Questions)** | | | |
| 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<br>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.<br>(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?<br>(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.<br>When at the lowest conceptuallevel<br>(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.<br>(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.<br>Write the following<br>(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.<br>Write the following<br>(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).<br>(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).<br>(b) For each technique, explain how each of the following functions may be implemented:<br>i. The generation of a data warehouse (including aggregation)<br>ii. Roll-up | Remember | CO 1 | AIT006.03 |

|  |  |  |  |  |
|---|---|---|---|---|
|  | iii. Drill-down<br>Which implementation techniques do you prefer, and why? |  |  |  |
| 9 | Briefly compare the following concepts. You may use an example to Describe your points.<br>(a) Snowflake schema, fact constellation, star schema model<br>(b) Data cleaning, data transformation, refresh | Remember | CO 1 | AIT006.03 |
| 10 | Design a star schema model for the following data.<br>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. To minimize the size of the fact table,dimension identifiers (e.g.,timekeyanditemkey) are system-generated identifiers. | Remember | CO 1 | AIT006.03 |
| **UNIT – II**<br>**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 | Ilustrate 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:<br>(a)Is it hype?<br>(b) Is it a simple transformation of Technology developed from Databases, statistics, and machine learning?<br>(c) Describe how the evolutions of database technology lead to datamining?<br>(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 handlingMissing 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<br>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<br>min-max normalization by setting min = 0 and max =1<br>z-scorenormalization | Understand | CO 2 | AIT006.07 |
| **Part – C (Problem Solving and Critical Thinking)** | | | | |
| 1 | Suppose that the data for analysis includes the attribute age.<br>The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20,20,<br>21, 22, 22, 25, 25,25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52,70.<br>Solve the following:<br>(a) Mean of the data?Median?<br>(b) mode of the data? Comment on the data's modality (i.e.bimodal, trimodal etc.).<br>(c) midrange of thedata? | Understand | CO 2 | AIT006.05 |
| 2 | Suppose that the data for analysis includes the attribute age.<br>The age values for the data tuples are (in increasing order)<br>13, 15, 16, 16, 19,20,20,21,22,22,25,2525, 25, 30, 33, 33, 35, 35, 35, 35, 36,<br>40, 45, 46, 52, 70.<br>Solve the following:<br>(a) Can you find (roughly) the first quartile (Q1) and the thirdquartile (Q3) of thedata?<br>(b) Give the five-number summary of thedata.<br>(c) Show a box plot of thedata.<br>(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.<br>(a) Use smoothing by bin means to smooth the above data, using a bindepth of3.<br>Illustrate your steps. Comment on the effect of this technique for the given data<br>(b) How might you determine outliers in thedata?<br>(c) What other methods are there for datasmoothing? | 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<br>age 23 23 27 27 39 41 47 49<br>%fat 9.5 26.57.8 17.8 31.4 25.9 27.4<br>27.2 31.2<br>age 52 54 54 56 57 58 58 60<br>%fat 34.6 42.528.8 33.4 30.2 34.1 32.9<br>41.2 35.7<br>Examine the following<br>(a) the mean, median and standard deviation of age and%fat.<br>(b) Draw the box plots for age and%fat.<br>(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,<br>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)<br>13, 15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,<br>40, 45, 46, 52, 70.<br>(a) Use min-max normalization to transform the value 35 for age on tothe range [0.0,1.0].<br>(b) Use z-score normalization to transform the value 35 for age, wherethe standard deviation of age is 12.94years.<br>(c) Use normalization by decimal scaling to transform the value 35 forage.<br>(d) Comment on which method you would prefer to use for the givendata, giving reasons as towhy. | Remember | CO 2 | AIT006.07 |
| | **UNIT – III**<br>**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 |

<table>
<tr><td>9.</td><td>State confidence and minimum confidence for strong association rule?</td><td>Understand</td><td>CO 3</td><td>AIT006.10</td></tr>
<tr><td>10.</td><td>List out the steps in association rule mining?</td><td>Remember</td><td>CO 3</td><td>AIT006.10</td></tr>
<tr><td>11.</td><td>Describe the two kinds of closure checking?</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>12.</td><td>Describe the five categories of pattern mining constraints?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>13.</td><td>List out the techniques of efficiency of Apriori algorithm?</td><td>Remember</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>14.</td><td>List out the drawbacks of Apriori technique?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td></td><td colspan="4"><b>Part – B (Long Answer Questions)</b></td></tr>
<tr><td>1</td><td>State the terms frequent item sets, closed item sets and association rules?</td><td>Remember</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>2</td><td>Illustrate which algorithm is an influential algorithm for mining frequent<br>Item sets for boolean association rules? Describe with an example?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>3</td><td>Describe the different techniques to improve the efficiency of Apriori?</td><td>Remember</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>4</td><td>Illustrate the FP-growth algorithmwith an example?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>5</td><td>Describe how to mine the frequent item sets using vertical data format?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>6</td><td>Illustrate about mining multilevel association rules from transaction databases in detail?</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>7</td><td>Describe how to mine the multidimensional association rules fromrelational databases and data warehouses?</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>8</td><td>Describe briefly about the different correlation measures in association analysis?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>9</td><td>Illustrate about constraint-based association mining?</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>10</td><td>Describe the Apriori algorithm with example?</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td></td><td><b>Part – C (Problem Solving and Critical Thinking Questions)</b></td><td></td><td></td><td></td></tr>
<tr><td>1</td><td>The following
<table>
<tr><th>T ID</th><th>List of item ID’s</th></tr>
<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>
</table>
(a)List out all frequent item sets using Apriori with Min support count 2<br>(b)List out all of the strong association rules</td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>3</td><td>Illustrate about frequent item set? Write the Apriori algorithm for frequent item set generation? Describe with an example</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>4</td><td>Illustrate about Market basket analysis with suitable example</td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
<tr><td>5</td><td>How can we mine multilevel Association rules efficiently using concept hierarchies? Illustrate with an A-priori algorithm for the given dataset below.
<table>
<tr><th>TID</th><th>List of items</th></tr>
<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>
</table></td><td>Understand</td><td>CO 3</td><td>AIT006.09</td></tr>
</table>

| | | | | |
|---|---|---|---|---|
| | **MID 11** | | | |
| | **Part - A (Short Answer Questions);** | | | |
| 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 |
| | **Part – B (Long Answer Questions)** | | 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 |
| | **Part – C (Problem Solving and Critical Thinking Questions)** | | | |

| | | | | |
|---|---|---|---|---|
| 1 | Describe support and confidence by using with following transactional data and find frequent item sets using FP growth tree. | 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 |

<table>
<tr><td>2</td><td>A database has five transactions. Let min sup = 60% and min conf = 80%.<br><table><tr><th>TID</th><th>items bought</th></tr><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></table><br>Examine the following<br>(d) Find all frequent item sets usingFP-growth.<br>(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.):</td><td>Remember</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>3</td><td>Compare Apriori and FP growth algorithms for frequent item set mining in transactional databases. Apply these algorithms to the following data:<br><table><tr><th>TID</th><th>LIST OF ITEMS</th></tr><tr><td>1</td><td>Bread, Milk, Sugar, TeaPowder, Cheese, Tomato</td></tr><tr><td>2</td><td>Onion, Tomato, Chillies, Sugar, Milk</td></tr><tr><td>3</td><td>Milk, Cake, Biscuits, Cheese, Onion</td></tr><tr><td>4</td><td>Chillies, Potato, Milk, Cake, Sugar, Bread</td></tr><tr><td>5</td><td>Bread, Jam, Mik, Butter, Chilles</td></tr><tr><td>6</td><td>Butter, Cheese, Paneer, Curd, Milk, Biscuits</td></tr><tr><td>7</td><td>Onion, Paneer, Chilies, Garlic, Milk</td></tr><tr><td>8</td><td>Bread, Jam, Cake, Biscuits, Tomato</td></tr></table></td><td>Remember</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>4</td><td>A database has six transactions. Let min-sup = 50% and min-conf =75%.<br>Find all frequent item sets using Apriori algorithm. List all the strong association rules<br><table><tr><th>TID</th><th>List of items</th></tr><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></table></td><td>Remember</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td>5</td><td>Can we design a method that mines the complete set of frequent item sets without candidate generation? If yes, Describe with following example<br><table><tr><th>TID</th><th>List of items</th></tr><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></table></td><td>Understand</td><td>CO 3</td><td>AIT006.08</td></tr>
<tr><td colspan="5"><b>UNIT-IV<br>CLASSIFICATION AND PREDICTION</b></td></tr>
<tr><td></td><td colspan="4"><b>Part – A (Short Answer Questions)</b></td></tr>
<tr><td>1</td><td>State classification?</td><td>Understand</td><td>CO 4</td><td>AIT006.14</td></tr>
</table>

|  |  |  |  |  |
|---|---|---|---|---|
| 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 |
|  | **Part – B (Long Answer Questions)** |  |  |  |
| 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:<br>(i) shapes of clusters that can be determined;<br>(ii)input parameters that must be specified;and (iii) limitations.<br>(a)k-means (b)k-medoids | Remember | CO 4 | AIT006.13 |

| | | | | |
|---|---|---|---|---|
| | **Part – C (Problem Solving and Critical Thinking Questions)** | | | |
| 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<br>(a) converting the decisiontree to rules and then pruning the resulting rules, or<br>(b) pruning the decision tree and then converting the pruned tree to rules.<br>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 treemethod?<br>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 samedata. | Understand | CO 4 | AIT006.12 |
| **UNIT-V<br>CLUSTERING** | | | | |
| | **Part - A (Short Answer Questions)** | | | |
| 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 |
| | **Part - B (Long Answer Questions)** | | | |
| 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<br>a) BIRCH<br>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:<br>(a) Calculate the mean absolute deviation ofage.<br>(b) Calculate the z-score for the first fourmeasurements. | | 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, distanced-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 |
| | **Part – C (Problem Solving and Critical Thinking Questions)** | | | |
| 1 | Given the following measurements for the variable age: 48, 12, 25, 42,<br>28,43,33,35, 56, 28, standardize the variable by the following: Calculate<br>(a) The mean absolute deviation of age. (b)The z-score for the first fourmeasurements. | Understand | CO5 | AIT006.15 |
| 2 | Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36,8):<br>Calculate<br>(a) The Euclidean distance between the two objects. (b)The Manhattan distance between the two objects.<br>(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) intothreeclusters.<br>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<br>distance function is Euclidean distance.<br>Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively.<br>Use the k-means algorithm to show only<br>The three cluster centers after the first round of executionand<br>The final threeclusters | Understand | CO5 | AIT006.15 |
| 4 | Describe why is it that BIRCH encounters difficulties in finding clustersof arbitrary shape but OPTICS does not? Can you proposesome<br>Modifications to BIRCH to help it find clusters of arbitraryshape? | 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.<br>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 |

<table>
<tr><td>8</td><td>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.<br><table>
<tr><td>Subject</td><td>A</td><td>B</td></tr>
<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>
</table></td><td>Understand</td><td>CO5</td><td>AIT006.19</td></tr>
<tr><td>9</td><td>Given the following measurements for the variable age: 29, 31, 25, 41,<br>27,43,33,35<br>56, 28, standardize the variable by the following:<br>Calculate<br>The mean absolute deviation ofage.<br>The z-score for the first threemeasurements.</td><td>Understand</td><td>CO5</td><td>AIT006.16</td></tr>
<tr><td>10</td><td>Given two objects represented by the tuples (21, 2, 41, 11) and (21, 1, 32,6):<br>Calculate<br>(a) The Euclidean distance between the two objects. (b)The Manhattan distance between the two objects.<br>(c) The Minkowski distance between the two objects, using p = 2.</td><td>Understand</td><td>CO5</td><td>AIT006.18</td></tr>
</table>

**Preparedby:**

Mr.Ch Suresh Kumar Raju , Assistant Professor **HOD, CSE**