MARRI LAXMAN REDDY

Institute of Technology & Management

(Approved by AICTE, New Delhi & Affiliated JNTU, Hyderabad)
Dundigal, Quthbullapur (M), Hyderabad-500043.

Computer science and Engineering Department

DATA MINING

LAB MANUAL

2020-21

DATA MINING LAB

LAB OBJECTIVES

Upon successful completion of this Lab the student will be able to:

1. Practical exposure on implementation of well known data mining Experiments.

2. Exposure to real life data sets for analysis and prediction.

3. Learning performance evaluation of data mining algorithms in a supervised and an unsupervised setting.

4. Handling a small data mining project for a given practical domain.

5. This lab course is intended to introduce data mining techniques including predictive, descriptive and visualization modeling and their effective use in discovering interesting hidden patterns in large volume of data generated by businesses, science, web, and other sources.

6. Focus is on the main process of data mining such as data preparation, classification, clustering, association analysis, and pattern evaluation.

DATA MINING LAB

LAB OUTCOMES:

Upon successful completion of this Lab the student will be able to:

1. The data mining process and important issues around data cleaning, preprocessing and integration.

2. The principle algorithms and techniques used in data mining, such as clustering, association mining, classification and prediction.

3. After undergoing the course students will be able to Synthesize the data mining fundamental concepts and techniques from multiple perspectives.

4. Develop skills and apply data mining tools for solving practical problems Advance relevant programming skills.

5. Gain experience and develop research skills by reading the data mining literature.

LIST OF LAB EXERCISES AS PER JNTU

Credit Risk Assessment

Description: The business of banks is making loans. Assessing the credit worthiness so an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a customer is good or bad A bank's business rules regarding loans must consider two opposing factors .on the one hand a bank wants to make as many loans as possible .Interest on these loans is the banks profit source .On the other hand .a bank cannot afford to make too many bad loans Too many bad loans could lead to the collapse of the bank. The bank's loan policy must involve a compromise: not too strict ,and not too lenient.

To do the assignment .you first and foremost need some knowledge about the world of credit you can acquire such knowledge in a number of ways

1. Knowledge Engineering .Find a loan officer who is willing to talk interview her and try to represent her knowledge In the form of production rules.

2. Books find some training manuals for loan officers or perhaps a suitable textbook on finance Translate this knowledge from text form to production rule form.

3. Common sense imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.

4. Case histories Find records of actual cases where competent loan officers correctly judged when and not to approve a loan application.

The German Credit Data:

Actual historical credit data is not always easy to come by because of confidentiality rules here is one such dataset consisting of 1000 actual cases collected in Germany credit dataset (original) Excel spread sheet version of the German credit data (Down load from web)

In spite of the fact that the data is German .you should probably make use of it for this assignment (Unless you really can consult a real loan officer)

A few notes on the German dataset

- DM stands for Deutsche Mark the unit of currency worth about 90 cents Canadian (but looks and acts like a quarter)
- Owns _telephone German phone rates are much higher than in Canada so fewer people own telephones
- Foreign worker There are millions of these in Germany (many from Turkey) it is very hard to get German citizenship if you were not born of German parents

- There are 20 attributes used in judging a loan applicant The goal is the classify the applicant into one of two categories ,good or bad

List of Experiments:

Experiment 1: List all the categorical (or nominal) attributes and the real-valued attributes separately.

Experiment 2: What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes.

Experiment 3: One type of model that you can create is a Decision Tree - train a Decision Tree using the complete dataset as the training data. Report the model obtained after training.

Experiment 4: Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly? (This is also called testing on the training set) Why do you think you cannot get 100 % training accuracy?

Experiment 5: Is testing on the training set as you did above a good idea? Why or Why not?
6. One approach for solving the problem encountered in the previous question is using cross-validation? Describe what cross-validation is briefly. Train a Decision Tree again using cross-validation and report your results. Does your accuracy increase/decrease? Why?

Experiment 6: Check to see if the data shows a bias against "foreign workers" (attribute 20),or "personal-status"(attribute 9). One way to do this (perhaps rather simple minded) is to remove these attributes from the dataset and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. To remove an attribute you can use the reprocess tab in Weka's GUI Explorer. Did removing these attributes have any significant effect? Discuss.

Experiment 7: Another question might be, do you really need to input so many attributes to get good results? Maybe only a few would do. For example, you could try just having attributes 2, 3, 5, 7, 10, 17 (and 21, the class attribute (naturally)). Tryout some combinations. (You had removed two attributes in problem 7.Remember to reload the arff data file to get all the attributes initially beforeyou start selecting the ones you want.)

Experiment 8: Sometimes, the cost of rejecting an applicant who actually has a good credit (case 1) might be higher than accepting an applicant who has bad credit (case 2). Instead of counting the misclassifications equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. You can do this by using a cost matrix in Weka. Train your Decision Tree again and report the Decision Tree and cross-validation results. Are they significantly different from results obtained in problem 6 (using equal cost)?

Experiment 9: Do you think it is a good idea to prefer simple decision trees instead of having long complex decision trees? How does the complexity of a Decision Tree relate to the bias of the model?

Experiment 10: You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning - Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross-validation (you can do this in Weka) and report the Decision Tree you obtain ? Also, report your accuracy using the pruned model. Does your accuracy increase ?

Experiment 11: (Extra Credit): How can you convert a Decision Trees into "if-then-else rules". Make up your own small Decision Tree consisting of 2-3 levels and convert it into a set of rules. There also exist different classifiers that output the model in the form of rules - one such classifier in Weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one ! Can you predict what attribute that might be in this dataset? OneR classifier uses a single attribute to make decisions (it chooses the attribute based on minimum error). Report the rule obtained by training a one R classifier. Rank the performance of j48, PART and oneR.

LIST OF ADDITIONAL EXPERIMENTS FOR THE SEMESTER

S. No	Name of the experiment
1	Experiment on k-means Data Clustering algorithms on weather data set.
2	Experiment on hierarchal Data Clustering algorithms on weather data set.
3.	Experiments on k-means Data Clustering algorithms on web site data set.

REFERNCE BOOKS

1. Andrew Moore's Data Mining Tutorial (On decision trees and cross validation).

2. Decision trees (source: TAN, MSU)

3. TOM Mitchell's book slides (on Conceptual learning and decision trees)

4. WEKA resources

- Introduction to WEKA
- Download WEKA
- WEKA tutorial
- ARFF format
- Using WEKA from command line

CONTENT OF LAB EXPERIMENTS

Experiment: 1

Objective: List all the categorical (or nominal) attributes and the real valued attributes separately.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Outcome: Student must kwon the types of data and their characteristics.

Algorithm / Procedure:

Input: German Data Set with 20 attributes

Attributes:-
1. checking_status
2. duration
3. credit history
4. purpose
5. credit amount
6. savings_status
7. employment duration
8. installment rate
9. personal status
10. debitors
11. residence_since
12. property
14. installment plans
15. housing
16. existing credits
17. job
18. num_dependents
19. telephone
20. foreign worker

Method:

1. For each attribute of Germen data set identify type of data and define data type, either numeric or string.
    a. If attribute is string type, find the values of attribute.

b. If the value is discrete, define attribute as nominal or categorical attribute. Otherwise, define attribute as string.

2. Repeat step 1 until end of all attributes in data set.
3. Display list of categorical and numerical valued attributes.

Output:

Categorical or Nomianal attributes:-

1. checking_status
2. credit history
3. purpose
4. savings_status
5. employment
6. personal status
7. debtors
8. property
9. installment plans
10. housing
11. job
12. telephone
13. foreign worker

Real valued attributes:-

1. duration
2. credit amount
3. credit amount
4. residence
5. age
6. existing credits
7. num_dependents

Outcome: We can identify the types of data in given data set.

Experiment: 2

Objective: What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Algorithm / Procedure:

Input: German Data Set with 20 attributes
Attributes:-
1. checking_status
2. duration
3. credit history
4. purpose
5. credit amount
6. savings_status
7. employment duration
8. installment rate
9. personal status
10. debitors
11. residence_since
12. property
14. installment plans
15. housing
16. existing credits
17. job
18. num_dependents
19. telephone
20. foreign worker

Method:

1. For each attribute of Germen data set,
    a. Analyze the values of attribute.
    b. Find attribute, which can be used for making decision on credit.
2. Form sample rules on selected attribute to classify the customer as good.
3. Form the sample rules on selected attribute to classify the customer as bad.

Output:

According to me the following attributes may be crucial in making the credit risk assessment.

1. Credit_history
2. Employment
3. Property_magnitude
4. job
5. duration
6. crdit_amount
7. installment
8. existing credit

Basing on the above attributes, we can make a decision whether to give credit or not.

Outcome: students are able to identify the attribute from data

Experiment: 3

Objective: One type of model that you can create is a Decision tree. Train a Decision tree using the complete data set as the training data. Report the model obtained after training.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Input: German data set with 20 attributes

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:

Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select training data set from test data option.
3. Start classification.

Output: The following model obtained after training the data set.

J48 pruned tree

```
------------------
checking_status = <0
| foreign_worker = yes
| | duration <= 11
| | | existing_credits <= 1
| | | | property_magnitude = real estate: good (8.0/1.0)
| | | | property_magnitude = life insurance
| | | | | own_telephone = none: bad (2.0)
| | | | | own_telephone = yes: good (4.0)
| | | | property_magnitude = car: good (2.0/1.0)
| | | | property_magnitude = no known property: bad (3.0)
```

| | | existing_credits > 1: good (14.0)
| | duration > 11
| | | job = unemp/unskilled non res: bad (5.0/1.0)
| | | job = unskilled resident
| | | | purpose = new car
| | | | | own_telephone = none: bad (10.0/2.0)
| | | | | own_telephone = yes: good (2.0)
| | | | purpose = used car: bad (1.0)
| | | | purpose = furniture/equipment
| | | | | employment = unemployed: good (0.0)
| | | | | employment = <1: bad (3.0)
| | | | | employment = 1<=X<4: good (4.0)
| | | | | employment = 4<=X<7: good (1.0)
| | | | | employment = >=7: good (2.0)
| | | | purpose = radio/tv
| | | | | existing_credits <= 1: bad (10.0/3.0)
| | | | | existing_credits > 1: good (2.0)
| | | | purpose = domestic appliance: bad (1.0)
| | | | purpose = repairs: bad (1.0)
| | | | purpose = education: bad (1.0)
| | | | purpose = vacation: bad (0.0)
| | | | purpose = retraining: good (1.0)
| | | | purpose = business: good (3.0)
| | | | purpose = other: good (1.0)
| | | job = skilled
| | | | other_parties = none
| | | | | duration <= 30
| | | | | | savings_status = <100
| | | | | | | credit_history = no credits/all paid: bad (8.0/1.0)
| | | | | | | credit_history = all paid: bad (6.0)
| | | | | | | credit_history = existing paid
| | | | | | | | own_telephone = none
| | | | | | | | | existing_credits <= 1
| | | | | | | | | | property_magnitude = real estate
| | | | | | | | | | | age <= 26: bad (5.0)
| | | | | | | | | | | age > 26: good (2.0)
| | | | | | | | | | property_magnitude = life insurance: bad (7.0/2.0)
| | | | | | | | | | property_magnitude = car
| | | | | | | | | | | credit_amount <= 1386: bad (3.0)
| | | | | | | | | | | credit_amount > 1386: good (11.0/1.0)
| | | | | | | | | | property_magnitude = no known property: good (2.0)
| | | | | | | | | existing_credits > 1: bad (3.0)
| | | | | | | | own_telephone = yes: bad (5.0)
| | | | | | | credit_history = delayed previously: bad (4.0)
| | | | | | | credit_history = critical/other existing credit: good (14.0/4.0)
| | | | | | savings_status = 100<=X<500

| | | | | | | credit_history = no credits/all paid: good (0.0)
| | | | | | | credit_history = all paid: good (1.0)
| | | | | | | credit_history = existing paid: bad (3.0)
| | | | | | | credit_history = delayed previously: good (0.0)
| | | | | | | credit_history = critical/other existing credit: good (2.0)
| | | | | | savings_status = 500<=X<1000: good (4.0/1.0)
| | | | | | savings_status = >=1000: good (4.0)
| | | | | | savings_status = no known savings
| | | | | | | existing_credits <= 1
| | | | | | | | own_telephone = none: bad (9.0/1.0)
| | | | | | | | own_telephone = yes: good (4.0/1.0)
| | | | | | | existing_credits > 1: good (2.0)
| | | | | duration > 30: bad (30.0/3.0)
| | | | other_parties = co applicant: bad (7.0/1.0)
| | | | other_parties = guarantor: good (12.0/3.0)
| | | job = high qualif/self emp/mgmt: good (30.0/8.0)
| foreign_worker = no: good (15.0/2.0)
checking_status = 0<=X<200
| credit_amount <= 9857
| | savings_status = <100
| | | other_parties = none
| | | | duration <= 42
| | | | | personal_status = male div/sep: bad (8.0/2.0)
| | | | | personal_status = female div/dep/mar
| | | | | | purpose = new car: bad (5.0/1.0)
| | | | | | purpose = used car: bad (1.0)
| | | | | | purpose = furniture/equipment
| | | | | | | duration <= 10: bad (3.0)
| | | | | | | duration > 10
| | | | | | | | duration <= 21: good (6.0/1.0)
| | | | | | | | duration > 21: bad (2.0)
| | | | | | purpose = radio/tv: good (8.0/2.0)
| | | | | | purpose = domestic appliance: good (0.0)
| | | | | | purpose = repairs: good (1.0)
| | | | | | purpose = education: good (4.0/2.0)
| | | | | | purpose = vacation: good (0.0)
| | | | | | purpose = retraining: good (0.0)
| | | | | | purpose = business
| | | | | | | residence_since <= 2: good (3.0)
| | | | | | | residence_since > 2: bad (2.0)
| | | | | | purpose = other: good (0.0)
| | | | | personal_status = male single: good (52.0/15.0)
| | | | | personal_status = male mar/wid
| | | | | | duration <= 10: good (6.0)
| | | | | | duration > 10: bad (10.0/3.0)
| | | | | personal_status = female single: good (0.0)

| | | | duration > 42: bad (7.0)
| | | other_parties = co applicant: good (2.0)
| | | other_parties = guarantor
| | | | purpose = new car: bad (2.0)
| | | | purpose = used car: good (0.0)
| | | | purpose = furniture/equipment: good (0.0)
| | | | purpose = radio/tv: good (18.0/1.0)
| | | | purpose = domestic appliance: good (0.0)
| | | | purpose = repairs: good (0.0)
| | | | purpose = education: good (0.0)
| | | | purpose = vacation: good (0.0)
| | | | purpose = retraining: good (0.0)
| | | | purpose = business: good (0.0)
| | | | purpose = other: good (0.0)
| | savings_status = 100<=X<500
| | | purpose = new car: bad (15.0/5.0)
| | | purpose = used car: good (3.0)
| | | purpose = furniture/equipment: bad (4.0/1.0)
| | | purpose = radio/tv: bad (8.0/2.0)
| | | purpose = domestic appliance: good (0.0)
| | | purpose = repairs: good (2.0)
| | | purpose = education: good (0.0)
| | | purpose = vacation: good (0.0)
| | | purpose = retraining: good (0.0)
| | | purpose = business
| | | | housing = rent
| | | | | existing_credits <= 1: good (2.0)
| | | | | existing_credits > 1: bad (2.0)
| | | | housing = own: good (6.0)
| | | | housing = for free: bad (1.0)
| | | purpose = other: good (1.0)
| | savings_status = 500<=X<1000: good (11.0/3.0)
| | savings_status = >=1000: good (13.0/3.0)
| | savings_status = no known savings: good (41.0/5.0)
| credit_amount > 9857: bad (20.0/3.0)
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)
Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.03 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances 855 85.5 %
Incorrectly Classified Instances 145 14.5 %
Kappa statistic 0.6251
Mean absolute error 0.2312

Root mean squared error 0.34
Relative absolute error 55.0377 %
Root relative squared error 74.2015 %
Total Number of Instances 1000

Experiment: 4

Objective:
Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly?(This is also called testing on the training set) why do you think can not get 100% training accuracy?

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Input: German data set with 20 attributes

Procedure:

Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select training data set from test data option.
3. Start classification.

Output:
The following model obtained after training the data set.
In the above model we trained complete dataset and we classified credit good/bad for each of the examples in the dataset.
For example:
IF
purpose=vacation THEN
credit=bad
ELSE

purpose=business THEN
Credit=good
In this way we classified each of the examples in the dataset.
We classified 85.5% of examples correctly and the remaining 14.5% of examples are incorrectly classified. We can't get 100% training accuracy because out of the 20 attributes, we have some unnecessary attributes which are also been analyzed and trained.
Due to this the accuracy is affected and hence we can't get 100% training accuracy.

Outcome:
If we used our above model trained on the complete dataset and classified credit as good/bad for each of the examples in that dataset. We can not get 100% training accuracy only **85.5%** of examples, we can classify correctly.

Experiment: 5

Objective: Is testing on the training set as you did above a good idea? Why or why not?

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Input: German Data set with 20 attributes

Procedure:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select training data set from test data option.
3. Start classification.

Output:

The following model obtained after training the data set.According to the rules, for the maximum accuracy, we have to take 2/3 of the dataset as training set and the remaining 1/3 as test set. But here in the above model we have taken complete dataset as training set which results only 85.5% accuracy. This is done for the analyzing and training of the unnecessary attributes which does not make a crucial role in credit risk assessment. And by this complexity is increasing and finally it leads to the minimum accuracy. If some part of the dataset is used as a training set and the remaining as test set then it leads to the accurate results and the time for computation will be less.This is why, we prefer not to take complete dataset as training set.

Outcome: It is not good idea by using 100% training data set.

Experiment: 6

Objective:
One approach for solving the problem encountered in the previous question is using cross- validation? Describe what is cross validation briefly. Train a decision tree again using cross validation and report your results. Does accuracy increase/decrease? Why?

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Outcome: Cross-Validation Definition: The classifier is evaluated by cross validation using the number of folds that are entered in the folds text field.

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select cross validataion with fold size 2, 5 and 10 from test data option.
3. Start classification.

In Classify Tab, Select cross-validation option and folds size is 2 then Press Start Button, next time change as folds size is 5 then press start, and next time change as folds size is 10 then press start.

Output:
The following model obtained after training the data set.
Fold Size – 2 output:
Fold Size – 5 output:

Fold Size – 10 output:

Cross validation:-

In k-fold cross-validation, the initial data are randomly portioned into 'k' mutually exclusive subsets or folds D1, D2, D3, ..., Dk. Each of approximately equal size. Training and testing is performed 'k' times. In iteration I, partition Di is reserved as the test set and the remaining partitions are collectively used to train the model. That is in the first iteration subsets D2, D3, ..., Dk collectively serve as the training set in order to obtain as first model. Which is tested on Di. The second trained on the subsets D1, D3, ..., Dk and test on the D2 and so on….

J48 pruned tree :-

```
------------------
checking_status = <0
| foreign_worker = yes
| | duration <= 11
| | | existing_credits <= 1
| | | | property_magnitude = real estate: good (8.0/1.0)
| | | | property_magnitude = life insurance
| | | | | own_telephone = none: bad (2.0)
| | | | | own_telephone = yes: good (4.0)
| | | | property_magnitude = car: good (2.0/1.0)
| | | | property_magnitude = no known property: bad (3.0)
| | | existing_credits > 1: good (14.0)
| | duration > 11
| | | job = unemp/unskilled non res: bad (5.0/1.0)
| | | job = unskilled resident
| | | | purpose = new car
| | | | | own_telephone = none: bad (10.0/2.0)
| | | | | own_telephone = yes: good (2.0)
| | | | purpose = used car: bad (1.0)
| | | | purpose = furniture/equipment
| | | | | employment = unemployed: good (0.0)
| | | | | employment = <1: bad (3.0)
| | | | | employment = 1<=X<4: good (4.0)
| | | | | employment = 4<=X<7: good (1.0)
| | | | | employment = >=7: good (2.0)
| | | | purpose = radio/tv
| | | | | existing_credits <= 1: bad (10.0/3.0)
| | | | | existing_credits > 1: good (2.0)
| | | | purpose = domestic appliance: bad (1.0)
| | | | purpose = repairs: bad (1.0)
| | | | purpose = education: bad (1.0)
| | | | purpose = vacation: bad (0.0)
| | | | purpose = retraining: good (1.0)
```

| | | | purpose = business: good (3.0)
| | | | purpose = other: good (1.0)
| | | job = skilled
| | | | other_parties = none
| | | | | duration <= 30
| | | | | | savings_status = <100
| | | | | | | credit_history = no credits/all paid: bad (8.0/1.0)
| | | | | | | credit_history = all paid: bad (6.0)
| | | | | | | credit_history = existing paid
| | | | | | | | own_telephone = none
| | | | | | | | | existing_credits <= 1
| | | | | | | | | | property_magnitude = real estate
| | | | | | | | | | | age <= 26: bad (5.0)
| | | | | | | | | | | age > 26: good (2.0)
| | | | | | | | | | property_magnitude = life insurance: bad (7.0/2.0)
| | | | | | | | | | property_magnitude = car
| | | | | | | | | | | credit_amount <= 1386: bad (3.0)
| | | | | | | | | | | credit_amount > 1386: good (11.0/1.0)
| | | | | | | | | | property_magnitude = no known property: good (2.0)
| | | | | | | | | existing_credits > 1: bad (3.0)
| | | | | | | | own_telephone = yes: bad (5.0)
| | | | | | | credit_history = delayed previously: bad (4.0)
| | | | | | | credit_history = critical/other existing credit: good (14.0/4.0)
| | | | | | savings_status = 100<=X<500
| | | | | | | credit_history = no credits/all paid: good (0.0)
| | | | | | | credit_history = all paid: good (1.0)
| | | | | | | credit_history = existing paid: bad (3.0)
| | | | | | | credit_history = delayed previously: good (0.0)
| | | | | | | credit_history = critical/other existing credit: good (2.0)
| | | | | | savings_status = 500<=X<1000: good (4.0/1.0)
| | | | | | savings_status = >=1000: good (4.0)
| | | | | | savings_status = no known savings
| | | | | | | existing_credits <= 1
| | | | | | | | own_telephone = none: bad (9.0/1.0)
| | | | | | | | own_telephone = yes: good (4.0/1.0)
| | | | | | | existing_credits > 1: good (2.0)
| | | | | duration > 30: bad (30.0/3.0)
| | | | other_parties = co applicant: bad (7.0/1.0)
| | | | other_parties = guarantor: good (12.0/3.0)
| | | job = high qualif/self emp/mgmt: good (30.0/8.0)
| foreign_worker = no: good (15.0/2.0)
checking_status = 0<=X<200
| credit_amount <= 9857
| | savings_status = <100
| | | other_parties = none
| | | | duration <= 42

| | | | | personal_status = male div/sep: bad (8.0/2.0)
| | | | | personal_status = female div/dep/mar
| | | | | | purpose = new car: bad (5.0/1.0)
| | | | | | purpose = used car: bad (1.0)
| | | | | | purpose = furniture/equipment
| | | | | | | duration <= 10: bad (3.0)
| | | | | | | duration > 10
| | | | | | | | duration <= 21: good (6.0/1.0)
| | | | | | | | duration > 21: bad (2.0)
| | | | | | purpose = radio/tv: good (8.0/2.0)
| | | | | | purpose = domestic appliance: good (0.0)
| | | | | | purpose = repairs: good (1.0)
| | | | | | purpose = education: good (4.0/2.0)
| | | | | | purpose = vacation: good (0.0)
| | | | | | purpose = retraining: good (0.0)
| | | | | | purpose = business
| | | | | | | residence_since <= 2: good (3.0)
| | | | | | | residence_since > 2: bad (2.0)
| | | | | | purpose = other: good (0.0)
| | | | | personal_status = male single: good (52.0/15.0)
| | | | | personal_status = male mar/wid
| | | | | | duration <= 10: good (6.0)
| | | | | | duration > 10: bad (10.0/3.0)
| | | | | personal_status = female single: good (0.0)
| | | | duration > 42: bad (7.0)
| | | other_parties = co applicant: good (2.0)
| | | other_parties = guarantor
| | | | purpose = new car: bad (2.0)
| | | | purpose = used car: good (0.0)
| | | | purpose = furniture/equipment: good (0.0)
| | | | purpose = radio/tv: good (18.0/1.0)
| | | | purpose = domestic appliance: good (0.0)
| | | | purpose = repairs: good (0.0)
| | | | purpose = education: good (0.0)
| | | | purpose = vacation: good (0.0)
| | | | purpose = retraining: good (0.0)
| | | | purpose = business: good (0.0)
| | | | purpose = other: good (0.0)
| | savings_status = 100<=X<500
| | | purpose = new car: bad (15.0/5.0)
| | | purpose = used car: good (3.0)
| | | purpose = furniture/equipment: bad (4.0/1.0)
| | | purpose = radio/tv: bad (8.0/2.0)
| | | purpose = domestic appliance: good (0.0)
| | | purpose = repairs: good (2.0)
| | | purpose = education: good (0.0)

| | | purpose = vacation: good (0.0)
| | | purpose = retraining: good (0.0)
| | | purpose = business
| | | | housing = rent
| | | | | existing_credits <= 1: good (2.0)
| | | | | existing_credits > 1: bad (2.0)
| | | | housing = own: good (6.0)
| | | | housing = for free: bad (1.0)
| | | purpose = other: good (1.0)
| | savings_status = 500<=X<1000: good (11.0/3.0)
| | savings_status = >=1000: good (13.0/3.0)
| | savings_status = no known savings: good (41.0/5.0)
| credit_amount > 9857: bad (20.0/3.0)
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)
Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.07 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 705 70.5 %
Incorrectly Classified Instances 295 29.5 %
Kappa statistic 0.2467
Mean absolute error 0.3467
Root mean squared error 0.4796
Relative absolute error 82.5233 %
Root relative squared error 104.6565 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.84 0.61 0.763 0.84 0.799 0.639 good
0.39 0.16 0.511 0.39 0.442 0.639 bad
Weighted Avg. 0.705 0.475 0.687 0.705 0.692 0.639
=== Confusion Matrix ===
a b <-- classified as
588 112 | a = good
183 117 | b = bad

Experiment: 7

Objective:

Check to see if the data shows a bias against "foreign workers" or "personal-status". One way to do this is to remove these attributes from the data set and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. Did removing these attributes have any significantly effect? Discuss.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:

Classification after removing “foreign worker” attribute.

Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. In preprocessor, select “foreign worker” attribute from attribute list and remove.
3. Select classifier tab, choose J48 decision tree and select training data set from test data option.
4. Start classification.

Output: The following model obtained after training the data set.

Procedure:

Classification after removing “personal status” attribute.

Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. In preprocessor, select “personal status” attribute from attribute list and remove.
3. Select classifier tab, choose J48 decision tree and select training data set from test data option.
4. Start classification.

Output:

The following model obtained after training the data set. This increase in accuracy is because thus two attributes are not much important in training and analyzing by removing this, the time has been reduced to some extent and then it results in increase in the accuracy. The decision tree which is created is very large compared to the decision tree which we have trained now. This is the main difference between these two decision trees.

Outcome: With this observation we have seen, when "Foreign_worker "attribute is removed from the Dataset, the accuracy is decreased. So this attribute is important for classification.

Experiment: 8

Objective:
Another question might be, do you really need to input so many attributes to get good results? May be only a few would do. For example, you could try just having attributes 2,3,5,7,10,17 and 21. Try out some combinations.(You had removed two attributes in problem 7. Remember to reload the arff data file to get all the attributes initially before you start selecting the ones you want.)

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Prerequisites:
Cross-Validation Definition: The classifier is evaluated by cross validation using the number of folds that are entered in the folds text field.

Pseudo code

In pseudo code, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:
Classification after removing 2nd attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set ARFF file in Weka Explorer.
2. In preprocessor, select 2nd attribute from attribute list and remove.
3. Select classifier tab, choose J48 decision tree and select training data set from test data option.
4. Start classification.

Output: The following model obtained after training the data set.

Procedure:
Classification after removing 3rd attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. In preprocessor, select 3rd attribute from attribute list and remove.
3. Select classifier tab, choose J48 decision tree and select training data set from test data option.
4. Start classification.

Output: The following model obtained after training the data set.

Procedure: Classification after removing 5th attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

5. Open German data set arff file in Weka Explorer.
6. In preprocessor, select 5th attribute from attribute list and remove.
7. Select classifier tab, choose J48 decision tree and select training data set from test data option.
8. Start classification.

Output: The following model obtained after training the data set.

Procedure:
Classification after removing 7th attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

5. Open German data set arff file in Weka Explorer.
6. In preprocessor, select 7th attribute from attribute list and remove.
7. Select classifier tab, choose J48 decision tree and select training data set from test data option.
8. Start classification.

Output: The following model obtained after training the data set.
Procedure:
Classification after removing 10th attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

9. Open German data set arff file in Weka Explorer.
10. In preprocessor, select 10th attribute from attribute list and remove.
11. Select classifier tab, choose J48 decision tree and select training data set from test data option.
12. Start classification.

Outcome: The following model obtained after training the data set.

Procedure:
Classification after removing 17th attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

9. Open German data set arff file in Weka Explorer.
10. In preprocessor, select 17th attribute from attribute list and remove.
11. Select classifier tab, choose J48 decision tree and select training data set from test data option.
12. Start classification.

Outcome: The following model obtained after training the data set.

Procedure:
Classification after removing 2nd attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

13. Open German data set arff file in Weka Explorer.
14. In preprocessor, select 2nd attribute from attribute list and remove.
15. Select classifier tab, choose J48 decision tree and select training data set from test data option.
16. Start classification.

Outcome: The following model obtained after training the data set.

Procedure:
Classification after removing 21st attribute:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

13. Open German data set arff file in Weka Explorer.
14. In preprocessor, select 21st attribute from attribute list and remove.
15. Select classifier tab, choose J48 decision tree and select training data set from test data option.
16. Start classification.

Output:

The following model obtained after training the data set.

```
=== Classifier model (full training set) ===
J48 pruned tree
------------------
credit_history = no credits/all paid: bad (40.0/15.0)
credit_history = all paid
```

| employment = unemployed
| | duration <= 36: bad (3.0)
| | duration > 36: good (2.0)
| employment = <1
| | duration <= 26: bad (7.0/1.0)
| | duration > 26: good (2.0)
| employment = 1<=X<4: good (15.0/6.0)
| employment = 4<=X<7: bad (10.0/4.0)
| employment = >=7
| | job = unemp/unskilled non res: bad (0.0)
| | job = unskilled resident: good (3.0)
| | job = skilled: bad (3.0)
| | job = high qualif/self emp/mgmt: bad (4.0)
credit_history = existing paid
| credit_amount <= 8648
| | duration <= 40: good (476.0/130.0)
| | duration > 40: bad (27.0/8.0)
| credit_amount > 8648: bad (27.0/7.0)
credit_history = delayed previously
| employment = unemployed
| | credit_amount <= 2186: bad (4.0/1.0)
| | credit_amount > 2186: good (2.0)
| employment = <1
| | duration <= 18: good (2.0)
| | duration > 18: bad (10.0/2.0)
| employment = 1<=X<4: good (33.0/6.0)
| employment = 4<=X<7
| | credit_amount <= 4530
| | | credit_amount <= 1680: good (3.0)
| | | credit_amount > 1680: bad (3.0)
| | credit_amount > 4530: good (11.0)
| employment = >=7
| | job = unemp/unskilled non res: good (0.0)
| | job = unskilled resident: good (2.0/1.0)
| | job = skilled: good (14.0/4.0)
| | job = high qualif/self emp/mgmt: bad (4.0/1.0)
credit_history = critical/other existing credit: good (293.0/50.0)
Number of Leaves : 27
Size of the tree : 40
Time taken to build model: 0.01 seconds
=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances 764 76.4 %
Incorrectly Classified Instances 236 23.6 %
Kappa statistic 0.3386
Mean absolute error 0.3488

Root mean squared error 0.4176
Relative absolute error 83.0049 %
Root relative squared error 91.1243 %
Total Number of Instances 1000
=== Classifier model (full training set) ===
J48 pruned tree
---
credit_history = no credits/all paid: bad (40.0/15.0)
credit_history = all paid
| employment = unemployed
| | duration <= 36: bad (3.0)
| | duration > 36: good (2.0)
| employment = <1
| | duration <= 26: bad (7.0/1.0)
| | duration > 26: good (2.0)
| employment = 1<=X<4: good (15.0/6.0)
| employment = 4<=X<7: bad (10.0/4.0)
| employment = >=7
| | job = unemp/unskilled non res: bad (0.0)
| | job = unskilled resident: good (3.0)
| | job = skilled: bad (3.0)
| | job = high qualif/self emp/mgmt: bad (4.0)
credit_history = existing paid
| credit_amount <= 8648
| | duration <= 40: good (476.0/130.0)
| | duration > 40: bad (27.0/8.0)
| credit_amount > 8648: bad (27.0/7.0)
credit_history = delayed previously
| employment = unemployed
| | credit_amount <= 2186: bad (4.0/1.0)
| | credit_amount > 2186: good (2.0)
| employment = <1
| | duration <= 18: good (2.0)
| | duration > 18: bad (10.0/2.0)
| employment = 1<=X<4: good (33.0/6.0)
| employment = 4<=X<7
| | credit_amount <= 4530
| | | credit_amount <= 1680: good (3.0)
| | | credit_amount > 1680: bad (3.0)
| | credit_amount > 4530: good (11.0)
| employment = >=7
| | job = unemp/unskilled non res: good (0.0)
| | job = unskilled resident: good (2.0/1.0)
| | job = skilled: good (14.0/4.0)
| | job = high qualif/self emp/mgmt: bad (4.0/1.0)
credit_history = critical/other existing credit: good (293.0/50.0)

Number of Leaves : 27
Size of the tree : 40
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 703 70.3 %
Incorrectly Classified Instances 297 29.7 %
Kappa statistic 0.1759
Mean absolute error 0.3862
Root mean squared error 0.4684
Relative absolute error 91.9029 %
Root relative squared error 102.2155 %
Total Number of Instances 1000

Outcome: With this observation we have seen, when 3rd attribute is removed from the Dataset, the accuracy (83%) is decreased. So this attribute is important for classification. when 2nd and 10th attributes are removed from the Dataset, the accuracy(84%) is same. So we can remove any one among them. when 7th and 17th attributes are removed from the Dataset, the accuracy(85%) is same. So we can remove any one among them. If we remove 5th and 21st attributes the accuracy is increased, so these attributes may not be needed for the classification.

Experiment: 9

Objective:
Sometimes, The cost of rejecting an applicant who actually has good credit might be higher than accepting an applicant who has bad credit. Instead of counting the misclassification equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. By using a cost matrix in weak. Train your decision tree and report the Decision Tree and cross validation results. Are they significantly different from results obtained in problem 6.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudocode, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka GUI Explorer.
2. In classify tab then press choose button in that Select J48 decision tree and select Use training data set option from test data option.
3. In classify tab press More options button then we get classifier evaluation options window in that select cost sensitive evaluation the press set button then we get Cost Matrix Editor.
4. Change classes as 2 then press Resize button. We get 2 by 2 Cost Matrix. In cost matrix (0,1) location change value as 5, we get modified cost matrix is as follows:0.0 5.0
   1.0 0.0
5. Then close the cost matrix editor, then press ok button.

Then press start button.

Output: The following model obtained after training the data set.

Outcome: With this observation we have seen that ,total 700 customers in that 669 classified as good customers and 31 misclassified as bad customers. In total 300cusotmers, 186 classified as bad customers and 114 misclassified as good customers.

Experiment: 10

Objective: Do you think it is a good idea to prefect simple decision trees instead of having long complex decision tress? How does the complexity of a Decision Tree relate to the bias of the model?

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code:

In pseudo code, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:
Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select training data set from test data option.
3. Start classification.

Output: The following model obtained after training the data set.

Outcome: It is Good idea to prefer simple Decision trees, instead of having complex Decision tree.

Experiment: 11

Objective: You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning. Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross validation and report the Decision Trees you obtain? Also Report your accuracy using the pruned model Does your Accuracy increase?

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudo code, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure:

We can make our decision tree simpler by pruning the nodes. Created a decision tree by using J48 Technique for the complete dataset as the training data in Weka Explorer.

1. Open German data set arff file in Weka Explorer.
2. Select classifier tab, choose J48 decision tree and select training data set from test data option. Beside Choose Button press on J48 –c 0.25 –M2 text, it displays Generic Object Editor. Select Reduced Error pruning property as True then press OK.
3. Start classification.

Output: The following model obtained after training the data set.

Outcome: By using pruned model, the accuracy decreased. Therefore by pruning the nodes we can make our decision tree simpler.

Experiment: 12

Objective: How can you convert a Decision Tree into "if-then-else rules". Make up your own small Decision Tree consisting 2-3 levels and convert into a set of rules. There also exist different classifiers that output the model in the form of rules. One such classifier in weka is rules. PART, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one ! Can you predict what attribute that might be in this data set? One R classifier uses a single attribute to make decisions(it chooses the attribute based on minimum error).Report the rule obtained by training a one R classifier. Rank the performance of j48,PART, one R.

Recommended Hardware / Software Requirements:

- Hardware Requirements: Intel Based desktop PC with minimum of 166 MHZ or faster processor with at least 64 MB RAM and 100 MB free disk space.
- Weka

Pseudo code

In pseudo code, the general algorithm for building decision trees is:

1. Check for base cases
2. For each attribute *a*
    1. Find the normalized information gain ratio from splitting on *a*
3. Let *a_best* be the attribute with the highest normalized information gain
4. Create a decision *node* that splits on *a_best*
5. Recurse on the sub lists obtained by splitting on *a_best*, and add those nodes as children of *node*

Procedure: In Weka GUI Explorer, Select Classify Tab, In that Select **Use Training set** option .There also exist different classifiers that output the model in the form of Rules. Such classifiers in weka are "PART" and "OneR" . Then go to Choose and select **Rules** in that select PART and press start Button.

Outcome: The following model obtained after training the data set.

EXPERIMENT -I
VIVA QUESTIONS

Question 1. What Is Data Mining?

Answer :

Data mining is a process of extracting hidden trends within a datawarehouse. For example an insurance dataware house can be used to mine data for the most high risk people to insure in a certain geographial area.

Question 2. Differentiate Between Data Mining And Data Warehousing?

Answer :

Data warehousing is merely extracting data from different sources, cleaning the data and storing it in the warehouse. Where as data mining aims to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc.

E.g. a data warehouse of a company stores all the relevant information of projects and employees. Using Data mining, one can use this data to generate different reports like profits generated etc.

Question 3. What Is Data Purging?

Answer :

The process of cleaning junk data is termed as data purging. Purging data would mean getting rid of unnecessary NULL values of columns. This usually happens when the size of the database gets too large.

Question 4. What Are Cubes?

Answer :

A data cube stores data in a summarized version which helps in a faster analysis of data. The data is stored in such a way that it allows reporting easily.

E.g. using a data cube A user may want to analyze weekly, monthly performance of an employee. Here, month and week could be considered as the dimensions of the cube.

Question 5. What Are Olap And Oltp?

Answer :

An IT system can be divided into Analytical Process and Transactional Process.

OLTP – categorized by short online transactions. The emphasis is query processing, maintaining data integration in multi-access environment.

OLAP – Low volumes of transactions are categorized by OLAP. Queries involve aggregation and very complex. Response time is an effectiveness measure and used widely in data mining techniques.

Question 6. What Are The Different Problems That "data Mining" Can Solve?

Answer :

• Data mining helps analysts in making faster business decisions which increases revenue with lower costs.
• Data mining helps to understand, explore and identify patterns of data.
• Data mining automates process of finding predictive information in large databases.
• Helps to identify previously hidden patterns.

Question 7. What Are Different Stages Of "data Mining"?

Answer :

Exploration: This stage involves preparation and collection of data. it also involves data cleaning, transformation. Based on size of data, different tools to analyze the data may be required. This stage helps to determine different variables of the data to determine their behavior.

Model building and validation: This stage involves choosing the best model based on their predictive performance. The model is then applied on the different data sets and compared for best performance. This stage is also called as pattern identification. This stage is a little complex because it involves choosing the best pattern to allow easy predictions.

Deployment: Based on model selected in previous stage, it is applied to the data sets. This is to generate predictions or estimates of the expected outcome.

Question 8. What Is Discrete And Continuous Data In Data Mining World?

Answer :

Discreet data can be considered as defined or finite data. E.g. Mobile numbers, gender. Continuous data can be considered as data which changes continuously and in an ordered fashion. E.g. age.

Question 9. What Is Model In Data Mining World?

Answer :

Models in Data mining help the different algorithms in decision making or pattern matching. The second stage of data mining involves considering various models and choosing the best one based on their predictive performance.

Question 10. How Does The Data Mining And Data Warehousing Work Together?

Answer :

Data warehousing can be used for analyzing the business needs by storing data in a meaningful form. Using Data mining, one can forecast the business needs. Data warehouse can act as a source of this forecasting.

Question 11. What Is A Decision Tree Algorithm?

Answer :

A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

Question 12. What Is Naive Bayes Algorithm?

Answer :

Naive Bayes Algorithm is used to generate mining models. These models help to identify relationships between input columns and the predictable columns. This algorithm can be used in the initial stage of exploration. The algorithm calculates the probability of every state of each input column given predictable columns possible states. After the model is made, the results can be used for exploration and making predictions.

Question 13. Explain Clustering Algorithm?

Answer :

Clustering algorithm is used to group sets of data with similar characteristics also called as clusters. These clusters help in making faster decisions, and exploring data. The algorithm first identifies relationships in a dataset following which it generates a series of clusters based on the relationships. The process of creating clusters is iterative. The algorithm redefines the groupings to create clusters that better represent the data.

Question 14. What Is Time Series Algorithm In Data Mining?

Answer :

Time series algorithm can be used to predict continuous values of data. Once the algorithm is skilled to predict a series of data, it can predict the outcome of other series. The algorithm generates a model that can predict trends based only on the original dataset. New data can also be added that automatically becomes a part of the trend analysis.

E.g. Performance one employee can influence or forecast the profit.

Question 15. Explain Association Algorithm In Data Mining?

Answer :

Association algorithm is used for recommendation engine that is based on a market based analysis. This engine suggests products to customers based on what they bought earlier. The model is built on a dataset containing identifiers. These identifiers are both for individual cases and for the items that cases contain. These groups of items in a data set are called as an item set. The algorithm traverses a data set to find items that appear in a case. MINIMUM_SUPPORT parameter is used any associated items that appear into an item set.

Question 16. What Is Sequence Clustering Algorithm?

Answer :

Sequence clustering algorithm collects similar or related paths, sequences of data containing events. The data represents a series of events or transitions between states in a dataset like a series of web clicks. The algorithm will examine all probabilities of transitions and measure the differences, or distances, between all the possible sequences in the data set. This helps it to determine which sequence can be the best for input for clustering.

E.g. Sequence clustering algorithm may help finding the path to store a product of “similar” nature in a retail ware house.

Question 17. Explain The Concepts And Capabilities Of Data Mining?

Answer :

Data mining is used to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc. it is more commonly used to transform large amount of data into a meaningful form. Data here can be facts, numbers or any real time information like sales figures, cost, meta data etc. Information would be the patterns and the relationships amongst the data that can provide information.

Question 18. Explain How To Work With The Data Mining Algorithms Included In Sql Server Data Mining?

Answer :

SQL Server data mining offers Data Mining Add-ins for office 2007 that allows discovering the patterns and relationships of the data. This also helps in an enhanced analysis. The Add-in called as Data Mining client for Excel is used to first prepare data, build, evaluate, manage and predict results.

Question 19. Explain How To Use Dmx-the Data Mining Query Language.

Answer :

Data mining extension is based on the syntax of SQL. It is based on relational concepts and mainly used to create and manage the data mining models. DMX comprises of two types of statements: Data definition and Data manipulation. Data definition is used to define or create new models, structures.

Example:
CREATE MINING SRUCTURE
CREATE MINING MODEL
Data manipulation is used to manage the existing models and structures.
Example:
INSERT INTO
SELECT FROM .CONTENT (DMX)

Question 20. Explain How To Mine An Olap Cube?

Answer :

A data mining extension can be used to slice the data the source cube in the order as discovered by data mining. When a cube is mined the case table is a dimension.
Question 21. What Are The Different Ways Of Moving Data/databases Between Servers And Databases In Sql Server?

Answer :

There are several ways of doing this. One can use any of the following options:
- BACKUP/RESTORE,
- Dettaching/attaching databases,
- Replication,
- DTS,
- BCP,
- logshipping,
- INSERT...SELECT,
- SELECT...INTO,
- creating INSERT scripts to generate data.

Question 22. What Are The Benefits Of User-defined Functions?

Answer :

a. Can be used in a number of places without restrictions as compared to stored procedures.
b. Code can be made less complex and easier to write.
c. Parameters can be passed to the function.
d. They can be used to create joins and also be sued in a select, where or case statement.
e. Simpler to invoke.

Question 23. Define Pre Pruning?

Answer :

A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples.

Question 24. What Are Interval Scaled Variables?

Answer :

Interval scaled variables are continuous measurements of linear scale. For example, height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance.

Question 25. What Is A Sting?

Answer :

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

Question 26. What Is A Dbscan?

Answer :

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

Question 27. Define Density Based Method?

Answer :

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

Question 28. Define Chameleon Method?

Answer :

Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster.

Question 29. What Do U Mean By Partitioning Method?

Answer :

In partitioning method a partitioning algorithm arranges all the objects into various partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

Question 30. Define Genetic Algorithm?

Answer :

Enables us to locate optimal binary string by processing an initial random population of binary strings by performing operations such as artificial mutation , crossover and selection.

Question 31. What Is Ods?

Answer :

1. ODS means Operational Data Store.
2. A collection of operation or bases data that is extracted from operation databases and standardized, cleansed, consolidated, transformed, and loaded into an enterprise data architecture. An ODS is used to support data mining of operational data, or as the store for base data that is summarized for a data warehouse. The ODS may also be used to audit the data warehouse to assure summarized and derived data is calculated properly. The ODS may further become the enterprise shared operational database, allowing operational systems that are being reengineered to use the ODS as there operation databases.

Question 32. What Is Spatial Data Mining?

Answer :

Spatial data mining is the application of data mining methods to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasises the importance of developing data driven inductive approaches to geographical analysis and modeling.

Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for applied GIS-based decision-making. Recently, the task of integrating these two technologies has become critical, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realise the huge potential of the information hidden there. Among those organizations are:

* offices requiring analysis or dissemination of geo-referenced statistical data
* public health services searching for explanations of disease clusters
* environmental agencies assessing the impact of changing land-use patterns on climate change
* geo-marketing companies doing customer segmentation based on spatial location.

Question 33. What Is Smoothing?

Answer :

Smoothing is an approach that is used to remove the nonsystematic behaviors found in time series. It usually takes the form of finding moving averages of attribute values. It is used to filter out noise and outliers.

Question 34. What Are The Advantages Data Mining Over Traditional Approaches?

Answer :

Data Mining is used for the estimation of future. For example if we take a company/business organization by using the concept of Data Mining we can predict the future of business interms of Revenue (or) Employees (or) Cutomers (or) Orders etc.

Traditional approches use simple algorithms for estimating the future. But it does not give accurate results when compared to Data Mining.

Question 35. What Is Model Based Method?

Answer :

For optimizing a fit between a given data set and a mathematical model based methods are used. This method uses an assumption that the data are distributed by probability distributions. There are two basic approaches in this method that are
1. Statistical Approach
2. Neural Network Approach.

Question 36. What Is An Index?

Answer :

Indexes of SQL Server are similar to the indexes in books. They help SQL Server retrieve the data quicker. Indexes are of two types. Clustered indexes and non-clustered indexes. Rows in the table are stored in the order of the clustered index key.
There can be only one clustered index per table.
Non-clustered indexes have their own storage separate from the table data storage.
Non-clustered indexes are stored as B-tree structures.
Leaf level nodes having the index key and it's row locater.
Question 37. Mention Some Of The Data Mining Techniques?

Answer :

Statistics
Machine learning
Decision Tree
Hidden markov models
Artificial Intelligence
Genetic Algorithm
Meta learning
Question 38. Define Binary Variables? And What Are The Two Types Of Binary Variables?

Answer :

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric

variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

Question 39. Explain The Issues Regarding Classification And Prediction?

Answer :

Preparing the data for classification and prediction:

Data cleaning
Relevance analysis
Data transformation
Comparing classification methods
Predictive accuracy
Speed
Robustness
Scalability
Interpretability
Question 40. What Are Non-additive Facts?

Answer :

Non-Additive: Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table.
Question 41. What Is Meteorological Data?

Answer :

Meteorology is the interdisciplinary scientific study of the atmosphere. It observes the changes in temperature, air pressure, moisture and wind direction. Usually, temperature, pressure, wind measurements and humidity are the variables that are measured by a thermometer, barometer, anemometer, and hygrometer, respectively. There are many methods of collecting data and Radar, Lidar, satellites are some of them.

Weather forecasts are made by collecting quantitative data about the current state of the atmosphere. The main issue arise in this prediction is, it involves high-dimensional characters. To overcome this issue, it is necessary to first analyze and simplify the data before proceeding with other analysis. Some data mining techniques are appropriate in this context.

Question 42. Define Descriptive Model?

Answer :

It is used to determine the patterns and relationships in a sample data.

Data mining tasks that belongs to descriptive model:

Clustering
Summarization
Association rules
Sequence discovery
Question 43. What Is A Star Schema?

Answer :

Star schema is a type of organising the tables such that we can retrieve the result from the database easily and fastly in the warehouse environment.Usually a star schema consists of one or more dimension tables around a fact table which looks like a star,so that it got its name.

Question 44. What Are The Steps Involved In Kdd Process?

Answer :

Data cleaning
Data Mining
Pattern Evaluation
Knowledge Presentation
Data Integration
Data Selection
Data Transformation
Question 45. What Is A Lookup Table?

Answer :

A lookUp table is the one which is used when updating a warehouse. When the lookup is placed on the target table (fact table / warehouse) based upon the primary key of the target, it just updates the table by allowing only new records or updated records based on the lookup condition.

Question 46. What Is Attribute Selection Measure?

Answer :

The information Gain measure is used to select the test attribute at each node in the decision tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

Question 47. Explain Statistical Perspective In Data Mining?

Answer :

Point estimation
Data summarization
Bayesian techniques
Hypothesis testing
Regression
Correlation
Question 48. Define Wave Cluster?

Answer :

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

Question 49. What Is Time Series Analysis?

Answer :

A time series is a set of attribute values over a period of time. Time Series Analysis may be viewed as finding patterns in the data and predicting future values.

Question 50. Explain Mining Single ?dimensional Boolean Associated Rules From Transactional Databases?

Answer :

The apriori algorithm: Finding frequent itemsets using candidate generation Mining frequent item sets without candidate generation.

EXPERIMENT -II
VIVA QUESTIONS

Question 1. What Is Meta Learning?

Answer :

Concept of combining the predictions made from multiple models of data mining and analyzing those predictions to formulate a new and previously unknown prediction.

Question 2. Describe Important Index Characteristics?

Answer :

The characteristics of the indexes are:
* They fasten the searching of a row.
* They are sorted by the Key values.
* They are small and contain only a small number of columns of the table.
* They refer for the appropriate block of the table with a key value.

Question 3. What Is The Use Of Regression?

Answer :

Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula.

Question 4. What Is Dimensional Modelling? Why Is It Important ?

Answer :

Dimensional Modelling is a design concept used by many data warehouse desginers to build thier data warehouse. In this design model all the data is stored in two types of tables - Facts table and Dimension table. Fact table contains the facts/measurements of the business and the dimension table contains the context of measuremnets ie, the dimensions on which the facts are calculated.

Question 5. What Is Unique Index?

Answer :

Unique index is the index that is applied to any column of unique value.
A unique index can also be applied to a group of columns.

Question 6. What Are The Foundations Of Data Mining?

Answer :

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:
* Massive data collection
* Powerful multiprocessor computers
* Data mining algorithms

Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996.1 In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods.

Question 7. What Snow Flake Schema?

Answer :

Snowflake Schema, each dimension has a primary dimension table, to which one or more additional dimensions can join. The primary dimension table is the only table that can join to the fact table.

Question 8. Differences Between Star And Snowflake Schemas?

Answer :

Star schema - all dimensions will be linked directly with a fat table.
Snow schema - dimensions maybe interlinked or may have one-to-many relationship with other tables.

Question 9. What Is Hierarchical Method?

Answer :

Hierarchical method groups all the objects into a tree of clusters that are arranged in a hierarchical order. This method works on bottom-up or top-down approaches.

Question10 . What Is Cure?

Answer :

Clustering Using Representatives is called as CURE. The clustering algorithms generally work on spherical and similar size clusters. CURE overcomes the problem of spherical and similar size cluster and is more robust with respect to outliers.

Question 11. What Is Etl?

Answer :

ETL stands for extraction, transformation and loading.

ETL provide developers with an interface for designing source-to-target mappings, ransformation and job control parameter.
*Extraction
Take data from an external source and move it to the warehouse pre-processor database.
*Transformation
Transform data task allows point-to-point generating, modifying and transforming data.
*Loading
Load data task adds records to a database table in a warehouse.

Question 12. Define Rollup And Cube?

Answer :

Custom rollup operators provide a simple way of controlling the process of rolling up a member to its parents values.The rollup uses the contents of the column as custom rollup operator for each member and is used to evaluate the value of the member's parents.

If a cube has multiple custom rollup formulas and custom rollup members, then the formulas are resolved in the order in which the dimensions have been added to the cube.

Question 13. What Are The Different Problems That "data Mining" Can Solve?

Answer :

*Data mining helps analysts in making faster business decisions which increases revenue with lower costs.

*Data mining helps to understand, explore and identify patterns of data.

*Data mining automates process of finding predictive information in large databases.

*Helps to identify previously hidden patterns.

Question 14. What Are Different Stages Of "data Mining"?

Answer :

Exploration: This stage involves preparation and collection of data. it also involves data cleaning, transformation. Based on size of data, different tools to analyze the data may be required. This stage helps to determine different variables of the data to determine their behavior.

Model building and validation: This stage involves choosing the best model based on their predictive performance. The model is then applied on the different data sets and compared for best performance. This stage is also called as pattern identification. This stage is a little complex because it involves choosing the best pattern to allow easy predictions.

Deployment: Based on model selected in previous stage, it is applied to the data sets. This is to generate predictions or estimates of the expected outcome.

Question 15. Explain How To Use Dmx-the Data Mining Query Language?

Answer :

Data mining extension is based on the syntax of SQL. It is based on relational concepts and mainly used to create and manage the data mining models. DMX comprises of two types of statements: Data definition and Data manipulation. Data definition is used to define or create new models, structures.

Example:
CREATE MINING SRUCTURE
CREATE MINING MODEL

Data manipulation is used to manage the existing models and structures.

Example:
INSERT INTO
SELECT FROM .CONTENT (DMX)

16Q: Explain in detail what is Data Purging?

Data purging is an important step in maintaining appropriate data in the database.
Basically deleting unnecessary data or rows which have NULL values from the database is nothing but data purging. So if there is a need to load fresh data into the database table we need to utilized database purging activity. This will clear all unnecessary data in the database and helps in maintaining clean and meaningful data.

Data purging is a process where junk data that exists in the database gets cleared out.

17Q: Explain in detail what does CUBE mean?

Cube is nothing but a data storage place where the data can be stored and makes it easier for the user to deal with his/her reporting tasks. It helps in expeditie data analysis process.
For example:
Let's say the data related to an employee is stored in the form of cube. If you are evaluating the user performance based on weekly, monthly basis then week and month are considered to be the dimensions of the cube.

18Q: What are the different problems that "Data Mining" can solve in general?

Data Mining is a very important process where it could be used to validate and screen the data how it is coming through and the process can be defined based on the data mining results. By doing these activities, the existing process can be modified.
They are widely used in the following industries :

1. Marketing
2. Advertising
3. Services
4. Artificial Intelligence
5. Government intelligence

By following the standard principles a lot of illegal activities can be identified and dealt with. As the internet has evolved a lot of loops holes also evolved at the same time.

19Q: Explain the difference between OLAP and OLTP?

OLTP:
1. OLTP stands for Online Transaction and Processing.
2. This is useful in the applications which involves in a lot of transactions and high volumes of data. This type of

applications are mainly observed in Banking sectors, Air ticketing etc. The architecture used in OLTP is Client server architecture. It actually supports the transactions cross network as well.

OLAP:

1. OLAP stands for Online Analytical Processing.

2. It is widely used in applications where we need to support business data where complex calculations happen. Most of the time, the data is in low volumes. As this is being multidimensional database, the user will have insight of how the data is coming through the various sources.

Check Out Artificial Intelligenc Tutorials

20Q: Explain the different stages of "Data Mining"?

They are three different stages in Data Mining, they are as follows:

1. Exploration
2. Model building and validation
3. Deployment

Exploration is a stage where a lot of activities revolve around preparation and collection of different data sets. So activities like cleaning, transformation are also included. Based on the data sets available , different tools are necessary to analyze the data.

Model Building and validation:

In this stage, the data sets is validated by applying different models where the data sets are compared for best performance. This particular step is called as pattern identification. This is a tedious process because the user has to identify which pattern is best suitable for easy predictions.

Deployment:

Based on the previous step, the best pattern is applied for the data sets and it is used to generate predictions and it helps in estimating expected outcomes.

21Q: Explain what is Discrete and continuous data concepts in Data Mining world?

Discrete data can be classified as a defined data or a finite data. That has a meaning to itself. For example: Mobile numbers, gender.

Continuous data is nothing but a data that continuous changes in an orderly fashion. The example for continuous data is "Age".

22Q: Explain what is MODEL in terms of Data Mining subject?

Model is an important factor in Data Mining activities, it defines and helps the algorithms in terms of making decisions and pattern matching. The second step to is that they evaluate different models that are available and select a best suitable model for the validating the data sets.

23Q: Explain what is Naive Bayes Algorithm?

The Naive Bayes Algorithm is widely used to generate mining models. These models are generally used to identify the relationship between the input columns and the predicated columns that are available. This algorithm is widely used during the initial stages of the explorations.

24Q: Explain in detail about Clustering Algorithm?

1. The clustering algorithm is actually used on groups of data sets are available with a common characteristics, they are called as clusters.

2. As the clusters are formed, it helps to make faster decisions and exporting the data is also fast.

3. First of all the algorithm identifies the relationships that are available in the dataset and based on that it generates clusters. The process of creating clusters is also repetitive.

25Q: Explain what is time series algorithm in data mining?

1. This algorithm is a perfect fit for type of data where the values changes continuous based on the time. For example : Age
2. If the algorithm is skilled and tuned to predict the data set, then it will be successfully keep a track of the continuous data and predict the right data.
3. This algorithm generates a specific model which is capable of predicting the future trends of the the data based on the real original data sets.
4. In between the process new data can also be added in part of trend analysis.

26Q: Explain in detail about association algorithm in Data mining?

This algorithm is mainly used in recommendation engine for a specific market based analysis.
So the input for this algorithm would be the products or items that are bought by a specific customer, based on that purchase a recommendation engine will predict the best suitable products for the customers.

27Q: What is sequence clustering algorithm?

As the name itself states that the data is collected at different points which occurs at sequence of events. The different data sets are analyzed based on the sequence of data sets that occur based on the events. The data sets are analyzed and then best possible data input will be determined for clustering.
Example:
A sequence clustering algorithm will help the organization to specific a particular path to introduce a new product which has similar characteristics in a retail warehouse.

28Q: What are the different concepts and capabilities of Data Mining?

So Data Mining is primarily responsible to understand and get meaningful data from the data sets that are stored in the database.
In terms of exploring the data in data mining is definitely helpful because it can be used in the following areas:

1. Reporting
2. Planning
3. Strategies
4. Meaningful Patterns etc.

A large amount of data is cleaned as per the requirement and can be transformed into a meaningful data which can be helpful for decision making at the executive level.

Data mining is really helpful with the following types of data:

1. Data sets which are in the form of sales figures
2. Forecast values for the business projection
3. Cost
4. Metadata etc

Based on the data analyzed, the information can be analyzed and appropriate relationships are defined.

29Q: What is the best way to work with data mining algorithms that are included in SQL Server data mining?

With the use of SQL Server data mining offers an add on for MS office 2007. This will help to identify and discover the relationships with the data. This data is helpful in future for enhanced analysis.
The add on is called as " Data Mining client for excel". With this the users will be able to first prepare data, build and further manage and evaluate the data where the final output will predicting results.

30Q: How to use DMX- the data mining query language in detail?

DMX consists of two types of statements in general.

Data definition and Data Manipulation.

Data Definition:

This is used to define and create new models and structures.

Data Manipulation:

As the name itself depicts, the data is manipulated based on the requirement.

The usage is explained in detail by picking up an example:

1. Create Mining Structure
2. Create Mining Model
3. Data Manipulation that is used in existing structures and models.

With the syntax, it is

INSERT INTO

SELECT FROM. CONTENT (DMX)

31Q: What are the different functions of data mining?

The different functions of data mining are as follows:

1. Characterization
2. Association and correlation analysis
3. Classification
4. Prediction
5. Cluster analysis
6. Evolution analysis
7. Sequence analysis

32Q: Explain in detail what is data aggregation and Generalization?

Data Aggregation:

As the name itself is self explanatory , the data is aggregated altogether where a cube can be constructed for data analysis purposes.

Generalization:

It is a process where low level data is replaced by high level concept so the data can be generalized and meaningful.

33Q: Explain in detail about In Learning and Inclassification:

In Learning:

This is a model which is primarily used to analyze a particular training data set and it has training data samples that are selected from a selected population.

In Classification:

This model is primarily used for providing an estimation for a particular class by selecting test samples randomly. The term classification is usually determined by identifying a known class for a specific unknown data.

34Q: Explain in detail what is Cluster Analysis?

The term cluster analysis is an important human activity which is widely used in different applications. To be specific, this type of analysis is used in market research, pattern recognition, data analysis and image processing.

35Q: Explain about data mining interface?

The data mining interface is usually used for improving the quality of the queries that are used.

The data mining Interface is nothing but the GUI form for data mining activities.

36Q: Why Tuning data warehouse is needed, explain in detail?

The main aspect of data warehouse is that the data evolves based on the time frame and it is difficult to predict the behaviour because of its ad hoc environment. The database tuning is much difficult in an OLTP environment because of its ad hoc and real time transaction loads. Due to its nature, the need to data warehouse tuning is necessary and it will change the way how the data is utilized based on the need.

37. What is SCD?

SCD is defined as slowly changing dimensions, and it applies to the cases where record changes over time.

38. What are the types of SCD?

There are three types of SCD and they are as follows:

SCD 1 – The new record replaces the original record

SCD 2 – A new record is added to the existing customer dimension table

SCD 3 – A original data is modified to include new data

39. What is BUS Schema?

BUS schema consists of suite of confirmed dimension and standardized definition if there is a fact tables.

40. What is Star Schema?

Star schema is nothing but a type of organizing the tables in such a way that result can be retrieved from the database quickly in the data warehouse environment.

41. What is Snowflake Schema?

Snowflake schema which has primary dimension table to which one or more dimensions can be joined. The primary dimension table is the only table that can be joined with the fact table.

42. What is a core dimension?

Core dimension is nothing but a Dimension table which is used as dedicated for single fact table or datamart.

43. What is called data cleaning?

Name itself implies that it is a self explanatory term. Cleaning of Orphan records, Data breaching business rules, Inconsistent data and missing information in a database.

44. What is Metadata?

Metadata is defined as data about the data. The metadata contains information like number of columns used, fix width and limited width, ordering of fields and data types of the fields.

45. What are loops in Datawarehousing?

In datawarehousing, loops are existing between the tables. If there is a loop between the tables, then the query generation will take more time and it creates ambiguity. It is advised to avoid loop between the tables.

46. Whether Dimension table can have numeric value?

Yes, dimension table can have numeric value as they are the descriptive elements of our business.

47. What is the definition of Cube in Datawarehousing?

Cubes are logical representation of multidimensional data. The edge of the cube has the dimension members,and the body of the cube contains the data values.

48. What is called Dimensional Modelling?

Dimensional Modeling is a concept which can be used by dataware house designers to build their own datawarehouse. This model can be stored in two types of tables – Facts and Dimension table.

Fact table has facts and measurements of the business and dimension table contains the context of measurements.

49. What are the types of Dimensional Modeling?

There are three types of Dimensional Modeling and they are as follows:

- Conceptual Modeling
- Logical Modeling
- Physical Modeling

50. What is surrogate key?

Surrogate key is nothing but a substitute for the natural primary key. It is set to be a unique identifier for each row that can be used for the primary key to a table.

EXPERIMENT -III
VIVA QUESTIONS

Q.1. What are foundations of data mining?

Generally, we use it for a long process of research and product development. Also, we can say this evolution was started when business data was first stored on computers. We can also navigate through their data in real time. Data Mining is also popular in the business community. As this is supported by three technologies that are now mature: Massive data collection, Powerful multiprocessor computers, and Data mining algorithms.

Read to know more about Data Mining

Q.2. What is the scope of data mining?

Automated prediction of trends and behaviours- We use to automate the process of finding predictive information in large databases. Also, questions that required extensive hands-on analysis can now be answered from the data. Moreover, targeted marketing is a typical example of predictive marketing. As we also use data mining on past promotional mailings.
Automated discovery of previously unknown patterns – As we use data mining tools to sweep through databases. Also, to identify previously hidden patterns in one step. Basically, there is a very good example of pattern discovery. As it is the analysis of retail sales data. Moreover, that is to identify unrelated products that are often purchased together.
Q.3 What are advantages of data mining?

Basically, to find probable defaulters, we use data mining in banks and financial institutions. Also, this is done based on past transactions, user behaviour and data patterns.

Generally, it helps advertisers to push the right advertisements to the internet. Also, it surfer on web pages based on machine learning algorithms. Moreover, this way data mining benefit both possible buyers as well as sellers of the various products.

Basically, the retail malls and grocery stores peoples used it. Also, it is to arrange and keep most sellable items in the most attentive positions.

Read more about data Mining Advantages

Q.4. What are the cons of data mining?

Security: The time at which users are online for various uses, must be important. They do not have security systems in place to protect us. As some of the data mining analytics use software. That is difficult to operate. Thus they require a user to have knowledge based training. The techniques of data mining are not 100% accurate. Hence, it may cause serious consequences in certain conditions.
Read more about data mining Disadvantages

Q.5 Name Data mining techniques?

a. Classification Analysis

b. Association Rule Learning

c. Anomaly or Outlier Detection

d. Clustering Analysis

e. Regression Analysis

f. Prediction

g. Sequential Patterns

h. Decision trees

Read more about Data Mining Techniques

Q.6. Give a brief introduction to data mining process?

Basically, data mining is the latest technology. Also, it is a process of discovering hidden valuable knowledge by analyzing a large amount of data. Moreover. we have to store that data in different databases. As data mining is a very important process. It becomes an advantage for various industries.
Read more about Data Mining Process

Q.7. Name types of data mining?

a. Data cleaning

b. Integration

c. Selection

d. Data transformation

e. Data mining

f. Pattern evaluation

g. Knowledge representation

Q.8. Name the steps used in data mining?

a. Business understanding

b. Data understanding

c. Data preparation

d. Modeling

e. Evaluation

f. Deployment

Q.9. Name areas of applications of data mining?

a. Data Mining Applications for Finance

b. Healthcare

c. Intelligence

d. Telecommunication

e. Energy

f. Retail

g. E-commerce

h. Supermarkets

i. Crime Agencies

j. Businesses Benefit from data mining

Read more applications of data mining

Q.10. What is required technological drivers in data mining?

Database size: Basically, as for maintaining and processing the huge amount of data, we need powerful systems.

Query Complexity: Generally, to analyze the complex and large number of queries, we need a more powerful system.

Data Mining Interview Questions Answers for Freshers – Q. 1,2,3,4,5,7,8,9

Data Mining Interview Questions Answers for Experience – Q. 6,10

Q.11. Give an introduction to data mining query language?

It was proposed by Han, Fu, Wang, et al. for the DBMiner data mining system. Although, it was based on the Structured Query Language. These query languages are designed to support ad hoc and interactive data mining. Also, it provides commands for specifying primitives. We can use DMQL to work with databases and data warehouses as well. We can also use it to define data mining tasks. Particularly we examine how to define data warehouses and data marts in DMQL.

Read more about data query language

Q.12. What is Syntax for Task-Relevant Data Specification?

The Syntax of DMQL for specifying task-relevant data −
use database database_name
or
use data warehouse data_warehouse_name
in relevance to att_or_dim_list
from relation(s)/cube(s) [where condition] order by order_list
group by grouping_list

Q.13. What is Syntax for Specifying the Kind of Knowledge?

Syntax for Characterization, Discrimination, Association, Classification, and Prediction.

Q.14. Explain Syntax for Interestingness Measures Specification?

Interestingness measures and thresholds can be specified by the user with the statement − with
<interest_measure_name> threshold = threshold_value

Q.15. Explain Syntax for Pattern Presentation and Visualization Specification?

Generally, we have a syntax, which allows users to specify the display of discovered patterns in one or more forms.
display as <result_form>

Q.16. Explain Data Mining Languages Standardization?

This will serve the following purposes −

Basically, it helps the systematic development of data mining solutions.
Also, improves interoperability among multiple data mining systems and functions.
Generally, it helps in promoting education and rapid learning.
Also, promotes the use of data mining systems in industry and society.
Q.17. Explain useful data mining queries?

First of all, it helps to apply the model to new data, to make single or multiple predictions.
Also, you can provide input values as parameters, or in a batch.
While it gets a statistical summary of the data used for training. Also, extract patterns and rule of the typical case representing a pattern in the model.
Also, helps in extracting regression formulas and other calculations that explain patterns.
Get the cases that fit a particular pattern.
Further, it retrieves details about individual cases used in the model.
Also, it includes data not used in the analysis. Moreover, it retrains a model by adding new data or perform cross-prediction.
Q.18. Give a brief introduction to data mining knowledge discovery?

Generally, most people don't differentiate data mining from knowledge discovery. While others view data mining as an essential step in the process of knowledge discovery.

Read more about Data Mining From Knowledge Discovery

Q.19. Explain steps involved in data mining knowledge process?

Data Cleaning −

Basically, in this step, the noise and inconsistent data are removed.

Data Integration −

Moreover, in this step, multiple data sources are combined.

Data Selection −

Furthermore, in this step, data relevant to the analysis task are retrieved from the database.

Data Transformation −

Basically, in this step, data is transformed into forms appropriate for mining. Also, by performing summary or aggregation operations.

Data Mining −

In this, intelligent methods are applied in order to extract data patterns.

Pattern Evaluation −

While, in this step, data patterns are evaluated.

Knowledge Presentation −

Generally, in this step, knowledge is represented

Q.20. What are issues in data mining?

A number of issues that need to be addressed by any serious data mining package

Uncertainty Handling

Dealing with Missing Values

Dealing with Noisy data

Efficiency of algorithms

Constraining Knowledge Discovered to only Useful

Incorporating Domain Knowledge

Size and Complexity of Data

Data Selection

Understandably of Discovered Knowledge: Consistency between Data and Discovered Knowledge.

Data Mining Interview Questions Answers for Freshers – Q. 11,16,17,18,19

Data Mining Interview Questions Answers for Experience – Q. 12,13,14,15,20

Q.21. What are major elements of data mining, explain?

Generally, helps in an extract, transform and load transaction data onto the data warehouse system.

While it stores and manages the data in a multidimensional database system.

Also, provide data access to business analysts and information technology professionals.

Generally, analyze the data by application software.

While, it shows the data in a useful format, such as a graph or table

Q.22. Name different level of analysis of data mining?

a. Artificial Neural Networks

b. Genetic algorithms

c. Nearest neighbor method

d. Rule induction

e Data visualization

Q.23. Name methods of classification methods?

a. Statistical Procedure Based Approach

b Machine Learning Based Approach

c. Neural Network

d. Classification Algorithms

e. ID3 Algorithm

f. C4.5 Algorithm

g. K Nearest Neighbors Algorithm

H. Naïve Bayes Algorithm

i. SVM Algorithm

J. ANN Algorithm

K. 48 Decision Trees

l. Support Vector Machines

M. SenseClusters (an adaptation of the K-means clustering algorithm)

Read more about Data Mining Classification

Q.24. Explain Statistical Procedure Based Approach?

Especially, there are two main phases present to work on classification. Also, it can be easily identified within the statistical community.

While, the second, “modern” phase concentrated on more flexible classes of models. Also, in which many of which attempt has to take. Moreover, it provides an estimate of the joint distribution of the feature within each class. Further, that can, in turn, provide a classification rule.

Generally, statistical procedures have to characterize by having a precise fundamental probability model and that is used to provides a probability of being in each class instead of just a classification.

Also, we can assume that the techniques will use by statisticians. Hence some human involvement has to assume with regard to variable selection.

Also, transformation and overall structuring of the problem.

Q.25. Explain Machine Learning Based Approach?

Generally, it covers automatic computing procedures. Also, it was based on logical or binary operations. Further, we use to learn a task from a series of examples.

Here, we have to focus on decision-tree approaches. Also, ss classification results come from a sequence of logical steps.

Also, its principle would allow us to deal with more general types of data including cases. While, the number and type of attributes may vary.

Q.26. Explain ID3 Algorithm?

Generally, the id3 calculation starts with the original set as the root hub. Also, on every cycle, it emphasizes through every unused attribute of the set and figures. Moreover, the entropy of attribute. Furthermore, at that point chooses the attribute. Also, it has the smallest entropy value.

Q.27. Name methods of clustering?

They are classified into the following categories −

Partitioning Method
Hierarchical Method
Density-based Method
Grid-Based Method
Model-Based Method
Constraint-based Method
Read more about Data Mining Clustering

Q.28. What do OLAP and OLTP stand for?

Basically, OLAP is an acronym for Online Analytical Processing and OLTP is an acronym for Online Transactional Processing.

Q.29. Define metadata?

Basically, metadata is simply defined as data about data. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

Q.30. List the types of OLAP server?

Basically, there are four types of OLAP servers, namely Relational OLAP, Multidimensional OLAP, Hybrid OLAP, and Specialized SQL Servers.

EXPERIMENT -IV
VIVA QUESTIONS

Q.1. What is Fact Table?

It contains the measurement of business processes. Also, foreign keys for the dimension tables.

Q.2. What does subject-oriented data warehouse signify?

Basically, we use it to shows that the data warehouse stores the information around a particular subject. Such as product, customer, sales, etc.

Q.3. Explain is data mining?

As we know data mining is a set of methods. Also, in addition, we should apply it to a large and complex database.

Read more about the introduction to data mining

Q.4. What is a history of data mining?

First of all, in 1960s statisticians used the terms "Data Fishing" or "Data Dredging". That was to refer what they considered the bad practice of analyzing data. Consequently, the term "Data Mining" appeared around 1990 in the database community.

Q.5. What are techniques used for data mining?

a. Artificial neural networks –

Generally, we use data mining in many ways. Just like in ANN it is used for non-linear predictive models. Also, it should be learning through training. Also, resemble biological neural networks in structure.

b. Decision trees-

Generally, Tree-shaped structures are used to represent sets of decisions. Also, for the classification of dataset rules are generated.

c. Genetic algorithms –

Basically, most of the genetic algorithms are present with the use of data mining. Also, these are a genetic combination, mutation, and natural selection for optimization techniques.

Read more about Data Mining Algorithms

Q.6. Why is Data Mining hot cake topic for this generation?

As we know that data mining is having spacious applications especially today. Hence, it is the young, hot and promising field for the present generation. Also, a good thing about this is that it has an attracted a great deal of attention in the information industry and in society.

Q.7. What are applications areas of data mining?

Weather forecasting.
E-commerce.
Self-driving cars.
Hazards of new medicine.
Space research.

Fraud detection.
Stock trade analysis.
Business forecasting.
Social networks.
Customers likelihood.

Q.8. Explain applications of data mining?

First of all, a credit card company use to leverage vast warehouse of customer transaction data. Also, we need to perform this to identify customers consequently.
There are too many analysis methods, we need to the manufacturer it for data mining. Then select promotional strategies that best reach their target customer segments.
Read more about detail data mining applications

Q.9. Explain areas where data mining has good effects?

Predict future trends, customer purchase habits
Help with decision making
Improve company revenue and lower costs
Market basket analysis

Q.10. Explain areas where data mining has bad effects?

User privacy/security
Amount of data is overwhelming
Great cost at implementation stage
Possible misuse of information
Possible inaccuracy of data
Read more about Disadvantages of Data Mining

Q.11. What is classification?

It seems like these are the examples, where the data analysis task is Classification − A bank loan officer wants to analyze the data in order to know which customer is risky or which are safe. A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

Read more about classification in detail

Q.12. Why is classification needed?

Most noteworthy, in today's world of "big data", a large database is becoming a norm. Just imagine there present a database with many terabytes. As facebook alone crunches 600 terabytes of new data every single day. Also, the primary challenge of big data is how to make sense of it. Moreover, the sheer volume is not the only problem. Also, big data need to be diverse, unstructured and fast changing. Consider audio and video data, social media posts, 3D data or geospatial data. Also, this kind of data is not easily categorized or organized. Further, to meet this challenge, a range of automatic methods for extracting information.

Q.13. What is Clustering?

Generally, a group of abstract objects into classes of similar objects is made. Although, we treat a cluster of data objects as one group. Also, while performing cluster analysis, we first partition the set of data into groups. As it was based on data similarity. Then we need to assign the labels to the groups. Moreover, a main advantage of over-classification is that it is adaptable to changes. Also, it helps single out useful features that distinguish different groups.

Q.14. What is Cluster Analysis?

Basically, finding groups of objects such that the objects in a group will be like one another. Also, it's different from the objects in other groups.

Q. 15. Explain the grid-based method?

Particularly, in a grid-based method, the objects together form a grid. Also, object space is quantized into a finite number of cells that form a grid structure.

Q.16. Explain the density-based method?

As it is based on the notion of density. The main idea behind this method is to continue growing the given cluster.

Q.17. Explain Model-based Method?

In this method, basically, a model is hypothesized for each cluster to find the best fit of data for a given model. Also, we use this method to locates the clusters by clustering the density function.

Q.18. Explain constraint-based method?

Basically, a constraint is referred to the user expectation. Also, it provides us with an interactive way of communication with the clustering process. Although, it can be specified by the user or the application need.

Q.19. Explain what is not cluster analysis?

Supervised classification – Have class label information

Simple segmentation – Dividing students into different registration groups, by the last name

Results of a query – Basically, groupings are a result of an external specification

Graph partitioning – Some mutual relevance and synergy, but areas are not identical

Q.20. Name some Data mining best books?

a. " Introduction to data mining" by Tan, Steinbach & Kumar (2006)

b. An Introduction to Statistical Learning: with Applications in R

c. Data Science for Business: What you need to know about data mining and data-analytic thinking

d. Modeling With Data

e. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners

f. Data Mining: Practical Machine Learning Tools and Techniques

g. Probabilistic Programming & Bayesian Methods for Hackers

Q.21. How can you deal with multi-source problems?

Basically, there are many ways to deal with multi-source problems. Also, we need to identify all records that are similar. We also need to put them together into a single record. Although, a record must contain the necessary attributes having no redundancy.

Q.22. Hierarchical Clustering Algorithm is known to be?

The algorithm which uses to combines and divides groups are already existing. That helps to create a hierarchical structure showcasing the manner in which these groups are merged or divided.

Q.23. Give an explanation of collaborative filtering?

It is just simply an algorithm. As we use it useful for creating a recommendation system. Also, it totally depends upon the behavioral data of the user.

Q.24. What is OLAP?

OLAP (Online Analytical Processing):

In the multidimensional model, we need to organize data into multiple dimensions. Although, each dimension contains multiple levels of abstraction defined by concept hierarchies. Also, OLAP provides a user-friendly environment for interactive data analysis.

Q.25. What is "data mining Interface"?

For better feedback to a user during the construction of a query, data mining interface is used in GUI form. Furthermore the GUI of data mining query improves the quality of the query.

Q.26. What is Business Intelligence?

Business Intelligence is also known as D.S.S – Decision support system which refers to the technologies, application, and practices for the collection, integration, and analysis of the business-related information or data. In addition it even helps to see the data on the information itself.

Q.27. What is Dimension Table?

Basically, dimension table is a table which contains attributes of measurements stored in fact tables. Also, this table consists of hierarchies, categories, and logic that can be used to traverse in nodes.

Q.28. Explain tiers in the tight-coupling data mining architecture?

First of all, we can define data layer as a database. This layer is an interface for all data sources.

While we use data mining application layer to retrieve data from a database. Some transformation routine has to perform here.

Front-end layer provides the intuitive and friendly user interface for end-user.

Q.29. What does subject-oriented data warehouse signify?

Basically, subject-oriented signifies that the data warehouse stores the information around a particular subject such as product, customer, sales, etc.

Q.30. Explain 48 Decision Trees?

Basically, a decision tree is a predictive machine-learning model. As it uses to decides the target value of a new sample. Also, the internal nodes of a decision tree denote the different attributes. Although, the branches between the nodes tell us the possible values.

EXPERIMENT -V
VIVA QUESTIONS

Q.1. Explain the types of data mining?

a. Data Cleaning

In this process data gets cleaned. As we know data in the real world is noisy, inconsistent and incomplete. It includes a number of techniques. Such as filling in the missing values, combined compute. The output of data cleaning process is adequately cleaned data.

b. Data Integration

In this process data in integrated from different data sources into one. As data lies in different formats in a different location. We can store data in a database, text files, spreadsheets, documents, data cubes, and so on. Although, we can say data integration is so complex, tricky and difficult task. That is because normally data doesn't match the different sources. We use metadata to reduce errors in the data integration process. Another issue faced is data redundancy.

In this case, same data might be available in different tables in the same database. Data integration tries to reduce redundancy to the maximum possible level. As without affecting the reliability of data.

c. Data Selection

This is the process by which data relevant to the analysis is retrieved from the database. As this process requires large volumes of historical data for analysis. So, usually, the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored.

Read more about data mining introduction

Q.2. Explain how data mining performed in detail?

a. Business understanding

First, we have to understand the requirements. Then have to find what are the business requirements. Next, the current situation has to access by finding out the different resources, assumptions. Also, by considering other important factors. Then, to achieve the business objectives we need to create data mining. Finally, we have to establish a new data mining plan to achieve both business and data mining goals. The plan should be as detailed as possible.

b. Data understanding

First, this phase starts with the collection of initial data. As in this, we have to collect data from available sources. As we have to collect data to get familiar with the data. Also, in order to make data collection, we need some activities that need to be performed. Such as data load and data integration. Next, the "gross" or "surface" properties of acquired data need to be examined and reported.

Then, we need to explore the data needs by tackling the data mining question. That can be addressed using querying, reporting, and visualization. Finally, have to examine the data quality by answering some important question. Such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

c. Data preparation

In this data, preparation process our 90% time consumed in our project. Also, it's outcome is the final data set. Once we identify the data sources, then we need to select, clean, construct and have to format in the desired form. The

data exploration task has to do with a greater depth. That need to be carry during this phase to notice the patterns. That is based on business understanding.

d. Modeling

First, we have to select modeling techniques that we need to use for the prepared dataset. Next, we have to generate test scenario to validate the quality and validity of the model. Then, by using modeling tools we have to prepare one or more models on the dataset. Finally, by involving these models need to be assessed involving stakeholders. That is to make sure that created models are met business initiatives.

e. Evaluation

Particularly, in this case, have to evaluate the result in the context of the business goal. In this phase, due to new patterns, new business requirements occurs. That patterns have to discover in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

f. Deployment

The information, which we gain through data mining process, we need to present it. The information has to represent in such a way that stakeholders can use it whenever they want it. Based on the business requirements, the deployment phase could be creating a report.
Also, as complex as a repeatable data mining process across the organization. In this plans for deployment, maintenance, have to be created for implementation. and also future supports. From the project point, the final report needs to summary the project experiences. And, review the project to see what need to improved created learned lessons. The CRISP-DM offers a uniform framework for experience documentation and guidelines. In addition, the CRISP-DM can apply in various industries with different types of data.

Read more about Data Mining Process

Q.3. Name some terminologies of data mining?

a. Cleaning (cleansing)

It is a process of preparing data for a data mining activity. Obvious data errors are detected and corrected and missing data is replaced.

b. Confusion matrix

We use this matrix shows that counts of the actual versus predicted class values. It shows not only how well the model predicts. But also presents the details needed to see exactly where things may have gone wrong. Consequent When an association between two variables is defined, the second item is called the consequent.

c. Continuous

Continuous data can have any value in an interval of real numbers. That is, the value does not have to be an integer. Continuous is the opposite of discrete or categorical.nbsp; Cross-validation A method of estimating the accuracy of a classification or regression model.
Read more about Data Mining Terminologies

Q.4. Explain applications of data mining?

a. For Finance

We have to Increase customer loyalty by collecting and analyzing customer behavior data. Also, one needs to help banks. That predict customer behavior and launch relevant services and products. Helps in Discovering hidden

correlations between various financial indicators. That need to detect suspicious activities with a high potential risk. Generally, it identifies fraudulent or non-fraudulent actions. As it done by collecting historical data. And then turning it into valid and useful information.

b. Data mining applications for Healthcare

Basically, it provides government, regulatory and competitor information that can fuel competitive advantage. Although, it supports the R&D process. And then go-to-market strategy with rapid access to information at every phase. Generally, it discovers the relationships between diseases and the effectiveness of treatments. That is to identify new drugs or to ensure that patients receive appropriate, timely care. Also, it supports healthcare insurers in detecting fraud and abuse.

c. Data mining applications for Intelligence

Generally, it reveals hidden data related to money laundering, narcotics trafficking, etc. Also, helps in Improving intrusion detection with a high focus on anomaly detection. And identify suspicious activity from a day one. Basically, convert text-based crime reports into word processing files. That can be used to support the crime-matching process.
Read more about data mining application

Q.5. What are pros of data mining?

a. Marketing / Retail Marketing companies use data mining to build models. That was based on historical data to predict who will respond to the new marketing campaigns. Such as direct mail, online marketing campaign etc. As a result, marketers have an approach to selling profitable products to targeted customers.

b. Finance / Banking As data mining provides financial institutions information about loan information and credit reporting. By building a model from historical customer's data can determine good and bad loans. Besides, it helps banks detect fraudulent credit card transactions. That is to protect credit card's owner.

c. Governments We use data mining in government agencies. That is by digging and analyzing records of the financial transaction. That is to build patterns that can detect money laundering.

d. Banking/Crediting As data mining is also used in financial institutions in areas. Such as credit reporting and loan information.

e. Law enforcement We use data mining in law enforcers to identify criminal suspects. Also, apprehending these criminals by examining trends in location. And also in other patterns of behaviors.

Read more about Advantages of Data Mining

Q.6. Explain data mining techniques?

a. Decision Trees

It's the most common technique, we use for data mining. As because of its simplest structure. The root of decision tree act as a condition. Each answer leads to specific data that help us to determine final decision based upon it.

b. Sequential Patterns

As we use this to discover regular events, similar patterns in transaction data. The historical data of customers helps us to identify the past transactions in a year. Clustering: Having similar characteristics clusters objects have to form, by using automatic method. We use clustering, to define classes. Then suitable objects have to place in each class.

c. Prediction

We use this method defines the relationship between independent and dependent instances.

d. Association

It is also known as relation technique. Also, in this, we have to recognize a pattern. That it is based upon the relationship of items in a single transaction. Also, we can suggest the technique for market basket analysis. That is to explore the products that customer frequently demands.

Read more about Data Mining Techniques

Q.7. Explain data mining architecture?

Data mining system contains too many components. That is a data source, data warehouse server, data mining engine, and knowledge base.

a) Data Sources

There are so many documents present. That is a database, data warehouse, World Wide Web (WWW). That are the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

b) Database or Data Warehouse Server

The database server contains the actual data that is ready to be processed. Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.

c) Data Mining Engine

In data mining system data mining engine is the core component. It consists of a number of modules. That we used to perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.

d) Pattern Evaluation Modules

This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value. Also, it interacts with the data mining engine. That's the main focus is to search towards interesting patterns.

e) Graphical User Interface

We use this interface to communicate between the user and the data mining system. Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process. When the user specifies a query, this module interacts with the data mining system. Thus, displays the result in an easily understandable manner.

f) Knowledge Base

In whole data mining process, the knowledge base is beneficial. We use it to guiding the search for the result patterns. The knowledge base might even contain user beliefs and data from user experiences. That can be useful in the process of data mining. The data mining engine might get inputs from the knowledge. That is the base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base. That is on a regular basis to get inputs and also to update it.

Read more about data mining architecture in details

Q.8. Explain in detail Syntax for Specifying the Kind of Knowledge?

a. Characterization

The syntax for characterization is −
mine characteristics [as pattern_name] analyze {measure(s) }
The analyze clause, specifies aggregate measures, such as count, sum, or count%

b. Discrimination

The syntax for Discrimination is
− mine comparison [as {pattern_name]}
For {target_class } where {t arget_condition }
{versus {contrast_class_i }
where {contrast_condition_i}}
analyze {measure(s) }

c. Association

The syntax for Association is−
mine associations [as {pattern_name}] {matching {metapattern} }

d. Classification

The syntax for Classification is −
mine classification [as pattern_name] analyze classifying_attribute_or_dimension

e. Prediction

The syntax for prediction is −
mine prediction [as pattern_name] analyze prediction_attribute_or_dimension
{set {attribute_or_dimension_i= value_i}}

Q.9. What are aspects of data mining?

a. Data Integration: First of all the data is collected and integrated from all the different sources.

b. Data Selection: Generally, we may not all the data we have collected in the first step. Also, in this step, we select only those data which we think useful for data mining.

c. Data Cleaning: Generally, the data we have collected is not clean. And may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

d. Data Transformation: Basically, the data even after cleaning is not ready for mining. Also, we need to transform them into forms appropriate for mining. Thus, the techniques used to do this are smoothing, aggregation, normalization etc.

e. Data Mining: As now in this step, we are ready to apply data mining techniques on the data. Basically, it is to discover the interesting patterns. Hence, clustering and association analysis are among the many different techniques present. Also, as we used for data mining.

f. Pattern Evaluation and Knowledge Presentation: Generally, this step includes visualization, transformation, removing redundant patterns from the patterns we generated.

g. Decisions / Use of Discovered Knowledge: As this step is beneficial to us. Also, it helps to use the knowledge acquired to take better decisions.

Q.10. Explain in detail different level of data analysis?

a. Artificial Neural Networks: Basically, there are present non-linear predictive models. That learn through training and resemble biological neural networks in structure.

b. Genetic algorithms: Generally, there are optimization techniques that use this process. Such as genetic combination, mutation. Also, natural selection in a design based on the concepts of natural evolution.

c. Nearest neighbor method: Basically, it is a technique which classifies each record in a dataset. Also, as it is based on a combination of the classes of the k record(s) most similar to it in a historical data set.

d. Rule induction: Generally, the extraction of useful if-then rules from data based on statistical significance.

e. Data visualization: Basically, the visual interpretation of complex relationships in multidimensional data. Also, we use graphics tools to illustrate data relationships.

Q.11. Explain syntax for soncept hierarchy specification?

We use the following syntax to specify concept hierarchies−
use hierarchy <hierarchy> for <attribute_or_dimension>

We use different syntaxes to define different types of hierarchies such as−
-schema hierarchies
define hierarchy time_hierarchy on date as [date,month quarter,year] –
set-grouping hierarchies
define hierarchy age_hierarchy for age on customer as
level1: {young, middle_aged, senior} < level0: all
level2: {20, …, 39} < level1: young
level3: {40, …, 59} < level1: middle_aged
level4: {60, …, 89} < level1: senior
-operation-derived hierarchies
define hierarchy age_hierarchy for age on customer
as {age_category(1), …, age_category(5)}
:= cluster(default, age, 5)< all(age)
-rule-based hierarchies
define hierarchy profit_margin_hierarchy on item as
level_1: low_profit_margin < level_0: all
if (price – cost)< $50
level_1: medium-profit_margin < level_0: all
if ((price – cost) > $50) and ((price – cost) ≤ $250))
level_1: high_profit_margin < level_0: all

Q.12. What is business intelligence?

Business Intelligence is also known as DSS – Decision support system which refers to the technologies, application, and practices for the collection, integration, and analysis of the business-related information or data. Even, it helps to see the data on the information itself.

Q.13. Give an explanation of collaborative filtering.

Collaborative filtering can be said to be a simple algorithm used for creating a recommendation system that depends on the behavioral data of the user.

Q.14. Explain how to work with the data mining algorithms included in SQL server data mining?

SQL Server data mining offers Data Mining Add-ins for office 2007 that allows discovering the patterns and relationships of the data. This also helps in an enhanced analysis. The Add-in called Data Mining Client for Excel is used to first prepare data, build, evaluate, manage and predict results.

Q.15. Explain the concepts and capabilities of data mining?
Data mining is used to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc. it is more commonly used to transform a large amount of data into a meaningful form. Data here can be facts, numbers or any real-time information like sales figures, cost, metadata etc. The information would be the patterns and the relationships amongst the data that can provide information.

Follow below link for more interview questions and answers

Q.16. How do the data mining and data warehousing work together?

Data warehousing can be used for analyzing the business needs by storing data in a meaningful form. Using Data mining, one can forecast the business needs. A data warehouse can act as a source of this forecasting.

Q.17. What is model in data mining World?

Models in Data mining help the different algorithms in decision making or pattern matching. The second stage of data mining involves considering various models and choosing the best one based on their predictive performance.

Q.18. What is discrete and continuous data in data mining world?

Discrete data can be considered as defined or finite data. E.g. Mobile numbers, gender. Continuous data can be considered as data which changes continuously and in an ordered fashion. E.g. age.

Q.19. What are the different problems that "data Mining" can solve?

Data mining helps analysts in making faster business decisions which increases revenue with lower costs.

Data mining helps to understand, explore and identify patterns of data.
We use data mining to automate the process of finding predictive information in large databases.
Also, helps to identify previously hidden patterns.

Q.20. What is data purging?

The process of cleaning junk data is termed as data purging. Purging data would mean getting rid of unnecessary NULL values of columns. This usually happens when the size of the database gets too large.

Q.21. Explain what is not cluster analysis?

Supervised classification – Have class label information
Simple segmentation – Dividing students into different registration groups, by the last name
Results of a query – Basically, groupings are a result of an external specification
Graph partitioning – Some mutual relevance and synergy, but areas are not identical.

Read more about Cluster analysis

Q.22. Explain partitioning method?

Partitioning Method Suppose we are given a database of 'n' objects. And the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups.

That must need to satisfy the following requirements − Each group contains at least one object. Each object must belong to exactly one group.
Points to remember − If we have a given number of partitions (say k). Then the partitioning method will create an initial partitioning. Further, it uses the iterative relocation technique. That is to improve the partitioning by moving objects from one group to other.

Q.23. What are requirements of clustering in data mining?

a. Scalability We need highly scalable clustering algorithms to deal with large databases.

b. Ability to deal with different kinds of attributes Algorithms should be capable to be applied to any kind of data. Such as interval-based data, categorical, and binary data.

c. Discovery of clusters with attribute shape The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures. That tend to find a spherical cluster of small sizes.

d. High dimensionality The clustering algorithm should not only be able to handle low-dimensional data. Although, need to handle the high dimensional space.

e. Ability to deal with noisy data Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

f. Interpretability The clustering results should be interpretable, comprehensible, and usable.

Q.24. What are applications of cluster analysis?

We use clustering analysis in different applications. Such as market research, pattern recognition, data analysis, and image processing.
Clustering can also help marketers discover distinct groups in their customer base. Moreover, they can characterize their customer groups based on the purchasing patterns.
Basically, in the field of biology, it can be used to derive plant and animal taxonomies. categorize genes with similar functionalities and gain insight into structures inherent to populations.
Clustering also helps in identification of areas. That are of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city. That is according to house type, value, and geographic location.
Clustering also helps in classifying documents on the web for information discovery Also, we use clustering in outlier detection applications. Such as detection of credit card fraud. As a data mining function, cluster analysis serves as a tool. That is to gain insight into the distribution of data. Also, need to observe characteristics of each cluster.

Q.25. Explain M. SenseClusters (an adaptation of the K-means clustering algorithm)

We have made use of SenseClusters to classify the email messages. That is into different user-define folders. SenseCluster available package of Perl programs. As it was developed at the University of Minnesota Duluth. That we use for automatic text and document classification. The advantage of SenseClusters is that it does not need any training data; It makes use of unsupervised learning methods to classify the available data. Now, particularly in this section will understand the K-means clustering algorithm. That has been use in SenseClusters. Clustering is the process in which we divide the available data. That instances into a given number of sub-groups. These sub-groups are clusters, and hence the name “Clustering”.

Q.26. Explain support vector machines?

Support Vector Machines are supervised learning methods. That used for classification, as well as regression. The advantage of this is that they can make use of certain kernels to transform the problem. Such that we can apply linear classification techniques to non-linear data. Applying the kernel equations. That arranges the data instances in

a way within the multi-dimensional space. That there is a hyperplane that separates data instances of one kind from those of another. The kernel equations may be any function. That transforms the non-separable data in one domain into another domain. In which the instances become separable. Kernel equations may be linear, quadratic, Gaussian, or anything else. That achieves this particular purpose.

Q.27. Explain ANN algorithm?

Artificial neural networks are types of computer architecture inspired by biological neural networks. They are used to approximate functions. That can depend on a large number of inputs and are generally unknown. They are presented as systems of interconnected "neurons". That can compute values from inputs. Also, they are capable of machine learning as well as pattern recognition. Due to their adaptive nature. An artificial neural network operates by creating connections between many different processing elements. That each corresponding to a single neuron in a biological brain. These neurons may be actually constructed or simulated by a digital computer system. Each neuron takes many input signal. That produces a single output signal that is sent as input to another neuron.

Read more about ANN Algorithm

Q.28. Explain S.V.M algorithm?

SVM has attracted a great deal of attention in the last decade. It also applied to various domains applications. SVMs are used for learning classification, regression or ranking function. SVM is based on statistical learning theory and structural risk minimization principle. And have the aim of determining the location of decision boundaries. We call it as a hyperplane. That produces the optimal separation of classes. Thereby creating the largest possible distance between the separating hyperplane. Further, the instance need to proof. That is to reduce an upper bound on the expected generalization error.
Read more about SVM Algorithm

For Freshers – Interview Question for Data Mining. Q-22,27,28

For Experienced – Interview Question for Data Mining. Q-23,24,25,26

Q.29. Explain Naïve Bayes algorithm?

The Naive Bayes Classifier technique is based on Bayesian theorem. Particularly, we use it when the dimensionality of the inputs is high. The Bayesian Classifier is capable of calculating the possible output. It is also possible to add new raw data at runtime and have a better probabilistic classifier. This classifiers considers the presence of a particular feature of a class. That is unrelated to the presence of any other feature when the class variable is given.

Q.30. Explain K Nearest Neighbors algorithm?

The closest neighbor rule distinguishes the classification of an unknown data point. That is on the basis of its closest neighbor whose class is already known. M. Cover and P. E. Hart purpose k nearest neighbor (KNN). In which nearest neighbor is computed on the basis of estimation of k. That indicates how many nearest neighbors are to be considered to characterize. It makes use of the more than one closest neighbor to determine the class. In which the given data point belongs to and so it is called as KNN. These data samples are needed to be in the memory at the run time.

Read more about Machine Learning Algorithms

Q.31. Explain neural network?

The field of Neural Networks has arisen from diverse sources. That is ranging from understanding and emulating the human brain to broader issues. That is of copying human abilities such as speech and can be used in various fields. Such as banking, in classification program to categorize data as intrusive or normal.
Generally, neural networks consist of layers of interconnected nodes. That each node producing a non-linear function of its input. And input to a node may come from other nodes or directly from the input data. Also, need to

identify some nodes with the output of the network. On the basis of this, there are different applications for neural networks present. That involve recognizing patterns and making simple decisions about them.

Read More about Neural Network

Q.32. Explain classification algorithms?

It is one of the Data Mining techniques. We use it to analyze a given data set and takes each instance of it. It assigns this instance to a particular class. Such that classification error will be least. Hence, we use this to extract models. That define important data classes within the given data set. Classification is a two-step process. During the first step, the model is created by applying classification algorithm. That is on training data set. Then in the second step, the extracted model is tested against a predefined test data set. That is to measure the model trained performance and accuracy. So classification is the process to assign class label from a data set whose class label is unknown.

Read more data mining interview questions and answers

Q.33. Explain tiers in the tight-coupling data mining architecture?

Data layer: We can define data layer as a database or data warehouse systems. This layer is an interface for all data sources. We store data mining results in the data layer. Thus, we can present to end-user in form of reports or another kind of visualization. We use data mining application layer is to retrieve data from a database. Some transformation routine have to perform here. That is to transform data into the desired format. Then we have to process data using various data mining algorithms. Front-end layer provides the intuitive and friendly user interface for end-user. Further, to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

Q.34. What are the required technological drivers in data mining?

Basically, data mining applications are present for all size machines. Such as mainframe, workstations, clouds, client, and server. Further, the size of enterprise applications varies from 10 Gb to 100 Tb. To deliver the applications exceeding 100 Tb, we prefer NCR systems. The technological drivers are as: Database size: As for maintaining and processing the huge amount of data, we need powerful systems. Query Complexity: To analyze the complex and large number of queries, we need more powerful system

Q.35. How do we categorize data mining systems?

As there are too many data mining systems available. Also, some systems are specific. That we need to dedicate to a given data source. Further, according to various criteria, data mining systems have to categorize.

a. Classification according to the type of data source mined According to the type of data handle, have to perform classification of data mining. Such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

b. Classification according to the data model drawn on
Generally, this classification is done on the basis of a data model. Such as relational database, object-oriented database, data warehouse, transactional, etc.

c. Classification according to the king of knowledge discovered
In this classification we perform this on the basis of the kind of knowledge. Such as characterization, discrimination, association, classification, clustering, etc.

d. Classification according to mining techniques used
As data mining systems employ are use to provide different techniques. According to the data analysis, we have to do this classification. Such as machine learning, neural networks, genetic algorithms, , etc.

EXPERIMENT -VI
VIVA QUESTIONS

1. What is Datawarehousing?

A Datawarehouse is the repository of a data and it is used for Management decision support system. Datawarehouse consists of wide variety of data that has high level of business conditions at a single point in time.

In single sentence, it is repository of integrated information which can be available for queries and analysis.

2. What is Business Intelligence?

Business Intelligence is also known as DSS – Decision support system which refers to the technologies, application and practices for the collection, integration and analysis of the business related information or data. Even, it helps to see the data on the information itself.

3. What is Dimension Table?

Dimension table is a table which contain attributes of measurements stored in fact tables. This table consists of hierarchies, categories and logic that can be used to traverse in nodes.

4. What is Fact Table?

Fact table contains the measurement of business processes, and it contains foreign keys for the dimension tables.

Example – If the business process is manufacturing of bricks

Average number of bricks produced by one person/machine – measure of the business process

5. What are the stages of Datawarehousing?

There are four stages of Datawarehousing:

Datawarehouse
Datawarehouse

Offline Operational Database
Offline Data Warehouse
Real Time Datawarehouse
Integrated Datawarehouse

6. What is Data Mining?

Data Mining is set to be a process of analyzing the data in different dimensions or perspectives and summarizing into a useful information. Can be queried and retrieved the data from database in their own format.

7. What is OLTP?

OLTP is abbreviated as On-Line Transaction Processing, and it is an application that modifies the data whenever it received and has large number of simultaneous users.

8. What is OLAP?

OLAP is abbreviated as Online Analytical Processing, and it is set to be a system which collects, manages, processes multi-dimensional data for analysis and management purposes.

9. What is the difference between OLTP and OLAP?

Following are the differences between OLTP and OLAP:

OLTP	OLAP
Data is from original data source	Data is from various data sources
Simple queries by users	Complex queries by system
Normalized small database	De-normalized Large Database
Fundamental business tasks	Multi-dimensional business tasks

10. What is ODS?

ODS is abbreviated as Operational Data Store and it is a repository of real time operational data rather than long term trend data.

11. What is the difference between View and Materialized View?

A view is nothing but a virtual table which takes the output of the query and it can be used in place of tables.

A materialized view is nothing but an indirect access to the table data by storing the results of a query in a separate schema.

12. What is ETL?

ETL is abbreviated as Extract, Transform and Load. ETL is a software which is used to reads the data from the specified data source and extracts a desired subset of data. Next, it transform the data using rules and lookup tables and convert it to a desired state.

Then, load function is used to load the resulting data to the target database.

13. What is VLDB?

VLDB is abbreviated as Very Large Database and its size is set to be more than one terabyte database. These are decision support systems which is used to server large number of users.

14. What is real-time datawarehousing?

Real-time datawarehousing captures the business data whenever it occurs. When there is business activity gets completed, that data will be available in the flow and become available for use instantly.

15. What are Aggregate tables?

Aggregate tables are the tables which contain the existing warehouse data which has been grouped to certain level of dimensions. It is easy to retrieve data from the aggregated tables than the original table which has more number of records.

This table reduces the load in the database server and increases the performance of the query.

16. What is factless fact tables?

A factless fact tables are the fact table which doesn't contain numeric fact column in the fact table.

17. How can we load the time dimension?

Time dimensions are usually loaded through all possible dates in a year and it can be done through a program. Here, 100 years can be represented with one row per day.

18. What are Non-additive facts?

Non-Addictive facts are said to be facts that cannot be summed up for any of the dimensions present in the fact table. If there are changes in the dimensions, same facts can be useful.

19. What is conformed fact?

Conformed fact is a table which can be used across multiple data marts in combined with the multiple fact tables.

20. What is Datamart?

A Datamart is a specialized version of Datawarehousing and it contains a snapshot of operational data that helps the business people to decide with the analysis of past trends and experiences. A data mart helps to emphasizes on easy access to relevant information.

21. What is Active Datawarehousing?

An active datawarehouse is a datawarehouse that enables decision makers within a company or organization to manage customer relationships effectively and efficiently.

22. What is the difference between Datawarehouse and OLAP?

Datawarehouse is a place where the whole data is stored for analyzing, but OLAP is used for analyzing the data, managing aggregations, information partitioning into minor level information.

23. What is ER Diagram?

ER diagram is abbreviated as Entity-Relationship diagram which illustrates the interrelationships between the entities in the database. This diagram shows the structure of each tables and the links between the tables.

24. What are the key columns in Fact and dimension tables?

Foreign keys of dimension tables are primary keys of entity tables. Foreign keys of fact tables are the primary keys of the dimension tables.

25. What is SCD?

SCD is defined as slowly changing dimensions, and it applies to the cases where record changes over time.

26. What are the types of SCD?

There are three types of SCD and they are as follows:

SCD 1 – The new record replaces the original record

SCD 2 – A new record is added to the existing customer dimension table

SCD 3 – A original data is modified to include new data

27. What is BUS Schema?
BUS schema consists of suite of confirmed dimension and standardized definition if there is a fact tables.

28. What is Star Schema?

Star schema is nothing but a type of organizing the tables in such a way that result can be retrieved from the database quickly in the data warehouse environment.

29. What is Snowflake Schema?

Snowflake schema which has primary dimension table to which one or more dimensions can be joined. The primary dimension table is the only table that can be joined with the fact table.

30. What is a core dimension?

Core dimension is nothing but a Dimension table which is used as dedicated for single fact table or datamart.

31. What is called data cleaning?

Name itself implies that it is a self explanatory term. Cleaning of Orphan records, Data breaching business rules, Inconsistent data and missing information in a database.

32. What is Metadata?

Metadata is defined as data about the data. The metadata contains information like number of columns used, fix width and limited width, ordering of fields and data types of the fields.

33. What are loops in Datawarehousing?

In datawarehousing, loops are existing between the tables. If there is a loop between the tables, then the query generation will take more time and it creates ambiguity. It is advised to avoid loop between the tables.

34. Whether Dimension table can have numeric value?

Yes, dimension table can have numeric value as they are the descriptive elements of our business.

35. What is the definition of Cube in Datawarehousing?

Cubes are logical representation of multidimensional data. The edge of the cube has the dimension members,and the body of the cube contains the data values.

36. What is called Dimensional Modelling?

Dimensional Modeling is a concept which can be used by dataware house designers to build their own datawarehouse. This model can be stored in two types of tables – Facts and Dimension table.

Fact table has facts and measurements of the business and dimension table contains the context of measurements.

37. What are the types of Dimensional Modeling?

There are three types of Dimensional Modeling and they are as follows:

Conceptual Modeling
Logical Modeling
Physical Modeling

38. What is surrogate key?

Surrogate key is nothing but a substitute for the natural primary key. It is set to be a unique identifier for each row that can be used for the primary key to a table.

39. What is the difference between ER Modeling and Dimensional Modeling?

ER modeling will have logical and physical model but Dimensional modeling will have only Physical model.

ER Modeling is used for normalizing the OLTP database design whereas Dimensional Modeling is used for de-normalizing the ROLAP and MOLAP design.

40. What are the steps to build the datawarehouse?

Following are the steps to be followed to build the datawaerhouse:

Gathering business requirements
Identifying the necessary sources
Identifying the facts
Defining the dimensions
Defining the attributes
Redefine the dimensions and attributes if required
Organize the Attribute hierarchy
Define Relationships
Assign unique Identifiers
41. What are the different types of datawarehosuing?

Following are the different types of Datawarehousing:

Enterprise Datawarehousing
Operational Data Store
Data Mart
42. What needs to be done while starting the database?

Following need to be done to start the database:

Start an Instance
Mount the database
Open the database
43. What needs to be done when the database is shutdown?

Following needs to be done when the database is shutdown:

Close the database
Dismount the database
Shutdown the Instance
44. Can we take backup when the database is opened?

No, We cannot take full backup when the database is opened.

45. What is defined as Partial Backup?

A Partial backup in an operating system is a backup short of full backup and it can be done while the database is opened or shutdown.

46. What is the goal of Optimizer?

The goal to Optimizer is to find the most efficient way to execute the SQL statements.

47. What is Execution Plan?

Execution Plan is a plan which is used to the optimizer to select the combination of the steps.

48. What are the approaches used by Optimizer during execution plan?

There are two approaches:

Rule Based
Cost Based
49. What are the tools available for ETL?

Following are the ETL tools available:

1. Informatica
2. Data Stage
3. Oracle
4. Warehouse Builder
5. Ab Initio
6. Data Junction

50.What is the difference between metadata and data dictionary?
Metadata is defined as data about the data. But, Data dictionary contain the information about the project information, graphs, abinito commands and server information.

EXPERIMENT -VII
VIVA QUESTIONS

1. What is data discrimination?

Data discrimination is the comparison of the general features of the target class objects against one or more contrasting objects.

2. What can business analysts gain from having a data warehouse?

First, having a data warehouse may provide a competitive advantage by presenting relevant information from which to measure performance and make critical adjustments in order to help win over competitors.

Second, a data warehouse can enhance business productivity because it is able to quickly and efficiently gather information that accurately describes the organization.

Third, a data warehouse facilitates customer relationship management because it provides a consistent view of customers and item across all lines of business, all departments and all markets.

Finally, a data warehouse may bring about cost reduction by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner.

3. Why is association rule necessary?

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases.

It is intended to identify strong rules discovered in database using different measures of interesting.

4. What are two types of data mining tasks?

Descriptive task

Predictive task

5. Define classification.

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

6. What are outliers?

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are called outliers.

7. What do you mean by evolution analysis?

Data evolution analysis describes and models regularities or trends for objects whose behavior change over time.

Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data.

Distinct features of such as analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

8. Define KDD.

The process of finding useful information and patterns in data.

9. What are the components of data mining?

Database, Data Warehouse, World Wide Web, or other information repository

Database or Data Warehouse Server

Knowledge Based

Data Mining Engine

Pattern Evaluation Module

User Interface

10. Define metadata.

A database that describes various aspects of data in the warehouse is called metadata.

11Q: Define data warehouse?

A : Data warehouse is a subject oriented, integrated, time-variant, and nonvolatile collection of data that supports management's decision-making process.

12Q: What does subject-oriented data warehouse signify?

A : Subject oriented signifies that the data warehouse stores the information around a particular subject such as product, customer, sales, etc.

13Q: List any five applications of data warehouse.

A : Some applications include financial services, banking services, customer goods, retail sectors, controlled manufacturing.

14Q: What do OLAP and OLTP stand for?

A : OLAP is an acronym for **Online Analytical Processing** and OLTP is an acronym of Online Transactional Processing.

Q: What is the very basic difference between data warehouse and operational databases?

15A : A data warehouse contains historical information that is made available for analysis of the business whereas an operational database contains current information that is required to run the business.

16Q: List the Schema that a data warehouse system can implements.

A : A data Warehouse can implement star schema, snowflake schema, and fact constellation schema.

17Q: What is Data Warehousing?

A : Data Warehousing is the process of constructing and using the data warehouse.

18Q: List the process that are involved in Data Warehousing.

A : Data Warehousing involves data cleaning, data integration and data consolidations.

19Q: List the functions of data warehouse tools and utilities.

A : The functions performed by Data warehouse tool and utilities are Data Extraction, Data Cleaning, Data Transformation, Data Loading and Refreshing.

20Q: What do you mean by Data Extraction?

A : Data extraction means gathering data from multiple heterogeneous sources.

21Q: Define metadata?

A : Metadata is simply defined as data about data. In other words, we can say that metadata is the summarized data that leads us to the detailed data.

22Q: What does Metadata Respiratory contain?

A : Metadata respiratory contains definition of data warehouse, business metadata, operational metadata, data for mapping from operational environment to data warehouse, and the algorithms for summarization.

23Q: How does a Data Cube help?

A : Data cube helps us to represent the data in multiple dimensions. The data cube is defined by dimensions and facts.

24Q: Define dimension?

A : The dimensions are the entities with respect to which an enterprise keeps the records.

25Q: Explain data mart.

A : Data mart contains the subset of organization-wide data. This subset of data is valuable to specific groups of an organization. In other words, we can say that a data mart contains data specific to a particular group.

26Q: What is Virtual Warehouse?

A : The view over an operational data warehouse is known as virtual warehouse.

27Q: List the phases involved in the data warehouse delivery process.

A : The stages are IT strategy, Education, Business Case Analysis, technical Blueprint, Build the version, History Load, Ad hoc query, Requirement Evolution, Automation, and Extending Scope.

28Q: Define load manager.

A : A load manager performs the operations required to extract and load the process. The size and complexity of load manager varies between specific solutions from data warehouse to data warehouse.

29Q: Define the functions of a load manager.

A : A load manager extracts data from the source system. Fast load the extracted data into temporary data store. Perform simple transformations into structure similar to the one in the data warehouse.

30Q: Define a warehouse manager.

A : Warehouse manager is responsible for the warehouse management process. The warehouse manager consist of third party system software, C programs and shell scripts. The size and complexity of warehouse manager varies between specific solutions.

31Q: Define the functions of a warehouse manager.

A : The warehouse manager performs consistency and referential integrity checks, creates the indexes, business views, partition views against the base data, transforms and merge the source data into the temporary store into the published data warehouse, backs up the data in the data warehouse, and archives the data that has reached the end of its captured life.

32Q: What is Summary Information?

A : Summary Information is the area in data warehouse where the predefined aggregations are kept.

33Q: What does the Query Manager responsible for?

A : Query Manager is responsible for directing the queries to the suitable tables.

34Q: List the types of OLAP server

A : There are four types of OLAP servers, namely Relational OLAP, Multidimensional OLAP, Hybrid OLAP, and Specialized SQL Servers.

35Q: Which one is faster, Multidimensional OLAP or Relational OLAP?

A : Multidimensional OLAP is faster than Relational OLAP.

36Q: List the functions performed by OLAP.

A : OLAP performs functions such as roll-up, drill-down, slice, dice, and pivot.

Q: How many dimensions are selected in Slice operation?

A : Only one dimension is selected for the slice operation.

37Q: How many dimensions are selected in dice operation?

A : For dice operation two or more dimensions are selected for a given cube.

38Q: How many fact tables are there in a star schema?

A : There is only one fact table in a star Schema.

39Q: What is Normalization?

A : Normalization splits up the data into additional tables.

40Q: Out of star schema and snowflake schema, whose dimension table is normalized?

A : Snowflake schema uses the concept of normalization.

41Q: What is the benefit of normalization?

A : Normalization helps in reducing data redundancy.

42Q: Which language is used for defining Schema Definition?

A : Data Mining Query Language (DMQL) is used for Schema Definition.

43Q: What language is the base of DMQL?

A : DMQL is based on Structured Query Language (SQL).

44Q: What are the reasons for partitioning?

A : Partitioning is done for various reasons such as easy management, to assist backup recovery, to enhance performance.

45Q: What kind of costs are involved in Data Marting?

A : Data Marting involves hardware & software cost, network access cost, and time cost.

EXPERIMENT -VIII
VIVA QUESTIONS

1.What is Data Mining? (100 % asked Data Mining Interview Questions)

Answer :

Data Mining is the process used for the extraction of hidden predictive data from huge databases.Everyone must be aware of data mining these days is an innovation also known as knowledge discovery process used for analyzing the different perspectives of data and encapsulate into proficient information.

Data Mining is process of discovering the patterns in very large data sets involving the different methods like Machine Learning,statistics,different database systems.

2.Define Data Mining. (100 % asked Data Mining Interview Questions)

Answer :

There are following different definitions of data mining :

Definition 1 :

Data Mining is the process used for the extraction of hidden predictive data from huge databases.

Definition 2 :

Data Mining is process of discovering the patterns in very large data sets involving the different methods like Machine Learning,statistics,different database systems.

Definition 3:

Data mining is defined as a process used to extract usable data from a larger set of any raw data which implies analyzing data patterns in large batches of data using one or more software.

Definition 4 :

The automated extraction of hidden data from a large amount of database is Data Mining.

Definition 5 :

Data mining refers to the process of extracting the valid and previously unknown information from a large database to make crucial business decisions.

3.What is basic difference between data mining and data warehousing?

Answer :

Data Warehousing :

Data warehousing is merely extracting data from different sources, cleaning the data and storing it in the warehouse.

Data Mining :

Where as data mining aims to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc.

Example :

A data warehouse of a company stores all the relevant information of projects and employees. Using Data mining, one can use this data to generate different reports like profits generated etc.

4.What are various features of Data Mining?

Answer:

Following are different features of data mining :

• Automatic pattern predictions based on trend and behaviour analysis.

• Prediction based on likely outcomes.

• Creation of decision-oriented information.

• Focus on large data sets and databases for analysis.

• Clustering based on finding and visually documented groups of facts not previously known.

5.Explain data purging?(100 % asked Data Mining Interview Questions)

Answer:

The process of cleaning junk data is termed as data purging. Purging data would mean getting rid of unnecessary NULL values of columns. This usually happens when the size of the database gets too large.

Data Purging is most important activity for database management systems. The junk data will grab the database memory and it will slows down the performance of the database. So frequent purging gives the fast performance of data.

6.Explain different types of Storage models in OLAP?(100 % asked Data Mining Interview Questions)

Answer :

The following are different types of storage models in OLAP :

1. MOLAP – Multidimensional Online Analytical Processing
2. ROLAP – Relational online Analytical processing
3. HOLAP – Hybrid online Analytical Processing

7.Explain MOLAP (Multidimensional Online Analytical Processing) with its Advantages and disadvantages?

Answer:

1. As the name itself depicts “MOLAP” , i.e. Multidimensional.

2. In this type of data storage, the data is stored in multidimensional cubes and not in the standard relational databases.

The advantage of using MOLAP is:

The query performance is excellent, this is because the data is stored in multidimensional cubes. Also the calculations are pre generated when a cube is created.

The disadvantage of using MOLAP is:

1. Only limited amount of data can be stored. Since the calculations are triggered at the cube generation process it cannot withstand huge amount of data.
2. Needs a lot of skill to utilize this.
3. Also it has licensing cost associated to it.

8.What Are Cubes?

Answer :

A data cube stores data in a summarized version which helps in a faster analysis of data. The data is stored in such a way that it allows reporting easily.

Example :

using a data cube A user may want to analyze weekly, monthly performance of an employee. Here, month and week could be considered as the dimensions of the cube.

9.What is OLAP ? Explain with example.

Answer:

OLAP is technology used in many Business Intelligence applications which includes complex analytical calculations.OLAP is used for complex calculations,Trends Analysis,sophisticated data modeling.OLAP database is stored in multidimensional database model.OLAP system contains less number of transactions but complex calculations like aggregation- Sum,count,average,min,max e.t.c.

The Aggregated data in OLAP system must be in months,quarters,years,weeks e.t.c. The key purpose to use OLAP system is to reduce the query response time and increase the effectiveness of reporting.If these aggregated calculations are already stored in repository and if user wants fast access of data then user can use OLAP system.OLAP database stores aggregated historical data in multidimensional schema.

Real Example :

If Company head wants information of Resources salary in year 2000.

In spite of using the transactional system we will use OLAP system here where aggregated data of year 2000 for Resources is already present.

10.What is OLTP transaction?Explain with example.

Answer:

OLTP system is known as large number of small daily transactions like insert,update and delete.Operational database is known as OLTP system.OLTP system provides fast query processing as well as it is also responsible to

provide data integrity and data consistency.The actual effectiveness of OLTP is measured in number of Transactions per second.OLTP normally contains current data and data normalization is used properly in OLTP system.

Real Example :

If Company head wants transactional report of all Employees In – Out time.

As Company head wants daily report of in-out time we need to provide it using OLTP system.We need to schedule report on daily basis using OLTP system.

11.What is difference between data warehouse and data mining?(100 % asked Data Mining Interview Questions)

Answer :

Data Warehousing:
It is a process where the data is extracted from various sources. Further, the data is cleansed and stored.

Data Mining:

1. It is a process where it explores the data using the queries.
2. Basically, the queries are used to explore a particular data set and examine the results. This will help the individual in reporting, strategy planning, visualizing meaningful data sets.

The above can be explained by taking a simple example:

1. Let's take a software company where all of their projects information is stored. This is nothing but Data Warehousing.
2. Accessing a particular project and identifying the Profit and Loss statement for that project can be considered as Data Mining.

12.What are different data mining techniques?

Answer:

Decision Trees: It's the most common technique used for data mining because of its simplest structure. The root of decision tree act as a condition or question with multiple answers. Each answer leads to specific data that help us to determine final decision based upon it.
Sequential Patterns: The pattern analysis used to discover regular events, similar patterns in transaction data. Like, in sales; the historical data of customers helps us to identify the past transactions in a year. Based on the historic purchasing frequency of customer, the best deals or offers have been introduced by business firms.
Clustering: Using the automatic method, cluster of objects is formed having similar characteristics. By using clustering, classes are defined and then suitable objects are placed in each class.
Prediction: This method discovers the relationship between independent and dependent instances. For example, in the area of sales; to predict the future profit, sale acts as independent instance and profit could be dependent. Then based on historical data of sales and profit, associated profit is predicted.
Association: Also called relation technique, in this a pattern is recognized based upon the relationship of items in a single transaction. It is suggested technique for market basket analysis to explore the products that customer frequently demands.

Classification: Based upon machine learning, used to classify each item in a particular set into predefined groups. This method adopts mathematical techniques such as neural networks, linear programming, and decision trees and so on.

13.What is ROLAP?Explain with advantages and disadvantages.

Answer :
As the name suggests that, the data is stored in the form of relational databases.

The advantages of using ROLAP is:

1. As the data is stored in relational databases, it can handle huge amount of data storage.
2. All the functionalities are available as this is a relational database.

The disadvantages of using ROLAP is:

1. It is comparatively slow.
2. All the limitations that apply to SQL , the same applies to ROLAP too.

14.Explain different usages of data mining.

Answer :

Following are some usages of data mining :

1.Fast Business Decisions :

Data mining helps analysts in making faster business decisions which increases revenue with lower costs.

2.Find Patterns:

Data mining helps to understand, explore and identify patterns of data.

3.Process automation:

Data mining automates process of finding predictive information in large databases.

4.Hidden Pattern Finding :

Helps to identify previously hidden patterns.

15.Tell different industries where data mining is frequently used?

Answer:

Following are different industries where data mining is frequently used :

1. Marketing
2. Advertising
3. Services
4. Artificial Intelligence

5. Government intelligence

16.Explain data mining examples. (at least 2 examples of data mining)

Answer :

The data mining is used in various industries.Following are two examples of data mining :

1.Mobile Service Providers :
The Mobile service providers uses huge data mining to collect customer data.Mobile phone and utilities companies use Data Mining and Business Intelligence to predict 'churn', the terms they use for when a customer leaves their company to get their phone/gas/broadband from another provider. They collate billing information, customer services interactions, website visits and other metrics to give each customer a probability score, then target offers and incentives to customers whom they perceive to be at a higher risk of churning.

2.Analytics Websites like Trivago :
There are different analytics websites which will compare the prices of different things from other website. The Analytics and data mining plays big role in that websites. If you check the website named Trivago which will gives the information of different hotel prices by comparing the different websites uses the predictive data mining technique which will mine the data from different websites and shows the results.

17.What are different stages of data mining ?(100 % asked Data Mining Interview Questions)

Answer :

There are following different stages of data mining :

a. Business understanding

b. Data understanding

c. Data preparation

d. Modeling

e. Evaluation

f. Deployment

18. Explain different stages of data mining?(100 % asked Data Mining Interview Questions)

Answer:

Stage 1 : Exploration

Exploration is a stage where a lot of activities revolve around preparation and collection of different data sets. So activities like cleaning, transformation are also included. Based on the data sets available , different tools are necessary to analyze the data.

Stage 2 : Model Building and validation
In this stage, the data sets is validated by applying different models where the data sets are compared for best performance. This particular step is called as pattern identification. This is a tedious process because the user has to identify which pattern is best suitable for easy predictions.

Stage 3 : Deployment:
Based on the previous step, the best pattern is applied for the data sets and it is used to generate predictions and it helps in estimating expected outcomes.

19.Explain Decision tree algorithm?

Answer :

A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

20.Why data warehouse tuning is needed? Explain.

Answer:

Performance tuning in data warehouse is needed because of its huge data.The data warehouse has very huge historical as well as current data.Its very difficult to fetch the specific pattern information within a specified time.The main aspect of data warehouse is that the data evolves based on the time frame and it is difficult to predict the behavior because of its ad hoc environment. The database tuning is much difficult in an OLTP environment because of its ad hoc and real time transaction loads. Due to its nature, the need to data warehouse tuning is necessary and it will change the way how the data is utilized based on the need.

21.What is cluster analysis in Data Mining?(100 % asked Data Mining Interview Questions)

Answer :

Clustering analysis is used to group sets of data with similar characteristics also called as clusters. These clusters help in making faster decisions, and exploring data. The algorithm first identifies relationships in a data-set following which it generates a series of clusters based on the relationships. The process of creating clusters is iterative. The algorithm redefines the groupings to create clusters that better represent the data.

22.How data warehouse and data mining work together?

Answer:

Following points gives you idea about the data warehouse and data mining relationship :

Data mining

1.Extracting useful information for large amounts of data, for the purpose of finding various methods for business intelligence. This is the process of data mining
2.Prediction of future is done by using data mining. Data warehousing is the source for data mining.

Data warehousing:

1.Extracting data from various resources, transforming into required form is done in data warehousing. Later this data is loaded into data warehouse.
2.Historical data is stored using data warehousing. Business analysis is done by business users.

23.What are different types of data mining?

Answer:

Following are different types of data mining :

a. Data cleaning

b. Integration

c. Selection

d. Data transformation

e. Data mining

f. Pattern evaluation

g. Knowledge representation

Data Mining Interview Questions

24.What is required technological drivers in data mining?

Answer:

Database size: Basically, as for maintaining and processing the huge amount of data, we need powerful systems.

Query Complexity: Generally, to analyze the complex and large number of queries, we need a more powerful system

25.What will be most common issues in data mining process ?

Answer:

There are following most common issues of data mining process :

A number of issues that need to be addressed by any serious data mining package

Uncertainty Handling

Dealing with Missing Values

Dealing with Noisy data

Efficiency of algorithms

Constraining Knowledge Discovered to only Useful

Incorporating Domain Knowledge

Size and Complexity of Data

Data Selection

Understandably of Discovered Knowledge: Consistency between Data and Discovered Knowledge.

26. Explain capabilities concept in Data Mining?

Answer :

Data mining is used to examine or explore the data using queries. These queries can be fired on the data warehouse. Explore the data in data mining helps in reporting, planning strategies, finding meaningful patterns etc. it is more commonly used to transform large amount of data into a meaningful form. Data here can be facts, numbers or any real time information like sales figures, cost, meta data etc. Information would be the patterns and the relationships amongst the data that can provide information.

27.Explain Data Aggregation and data Generalization?

Answer:

Data Aggregation:
As the name itself is self explanatory , the data is aggregated altogether where a cube can be constructed for data analysis purposes.

Generalization:
It is a process where low level data is replaced by high level concept so the data can be generalized and meaningful.

28.What are different level of analysis in data mining?

Answer:

a.Artificial neural network

b. Genetic algorithms

c. Nearest neighbor method

d. Rule induction

e Data visualization

29.What is machine learning ?

Answer:

Generally, it covers automatic computing procedures. Also, it was based on logical or binary operations. Further, we use to learn a task from a series of examples.

Here, we have to focus on decision-tree approaches. Also, ss classification results come from a sequence of logical steps.

Also, its principle would allow us to deal with more general types of data including cases. While, the number and type of attributes may vary.

30.What is Sting?

Answer :Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

EXPERIMENT -IX
VIVA QUESTIONS

1. Differentiate between star schema and snowflake schema.

•Star Schema is a multi-dimension model where each of its disjoint dimension is represented in single table.
•Snow-flake is normalized multi-dimension schema when each of disjoint dimension is represent in multiple tables.
•Star schema can become a snow-flake
•Both star and snowflake schemas are dimensional models; the difference is in their physical implementations.
•Snowflake schemas support ease of dimension maintenance because they are more normalized.
•Star schemas are easier for direct user access and often support simpler and more efficient queries.
•It may be better to create a star version of the snowflaked dimension for presentation to the users

02. List the advantages of star schema.

•Star Schema is very easy to understand, even for non technical business manager.
•Star Schema provides better performance and smaller query times
•Star Schema is easily extensible and will handle future changes easily

03. What are the characteristics of data warehouse?

Integrated
Non-volatile
Subject oriented
Time varient

04. Define support and confidence.

The support for a rule R is the ratio of the number of occurrences of R, given all occurrences of all rules.
The confidence of a rule X->Y, is the ratio of the number of occurrences of Y given X, among all other occurrences given X

05. What are the criteria on the basic of which classification and prediction can be compared?

speed, accuracy, robustness, scalability, goodness of rules, interpret-ability

06. What is Data purging?

The process of cleaning junk data is termed as data purging. Purging data would mean getting rid of unnecessary NULL values of columns. This usually happens when the size of the database gets too large.

07 .What is a source qualifier?

When you add a relational or a flat file source definition to a mapping, you need to connect it to a Source Qualifier transformation. The Source Qualifier represents the rows that the Informatica Server reads when it executes a session.

08. What are the steps involved in Database Startup?

Start an instance, Mount the Database and Open the Database.

09. What are data modeling and data mining? Where it will be used?

Data modeling is the process of designing a data base model. In this data model data will be stored in two types of table fact table and dimension table Fact table contains the transaction data and dimension table contains the master data. Data mining is process of finding the hidden trends is called the data mining.

10. What is a 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.

11. Which of the following forms the logical subset of the complete data warehouse?

(a)Dimensional model(b)Fact table(c)Dimensional table(d)Operational Data Store(e)Data Mart.

12.Which of the following is not included in Modeling Applications?

(a)Forecasting models(b)Behavior scoring models(c)Allocation models(d)Data mining Models(e)Metadata driven models.

13.Which of the following is a dimension that means the same thing with every possible fact table towhich it can be joined?

(a)Permissible snowflaking(b)Confirmed Dimensions(c)Degenerate dimensions(d)Junk Dimensions(e)Monster Dimensions.

14.Which of the following is not the managing issue in the modeling process?

(a)Content of primary units column(b)Document each candidate data source(c)Do regions report to zones(d)Walk through business scenarios(e)Ensure that the transaction edit flat is used for analysis.

15.Which of the following criteria is not used for selecting the data sources?

(a)Data Accessibility(b)Platform(c)Data accuracy

(d)Longevity of the feed(e)Project scheduling.

16.Which of the following does not relate to the data modeling tool?

(a)Link to the dimension table designs(b)Business user Documentation(c)Helps assure consistency in naming(d)Length of the logical column.(e)Generates physical object DDL.

17.Which of the following is true on building a Matrix for Data warehouse bus architecture?

(a)Data marts as columns and dimensions as rows(b)Dimensions as rows and facts as columns(c)Data marts as rows and dimensions as columns(d)Data marts as rows and facts as columns(e)Facts as rows and data marts as columns.

18.Which of the following should not be considered for each dimension attribute?

(a)Attribute name(b)Rapid changing dimension policy(c)Attribute definition(d)Sample data(e)Cardinality.

19.Which of following form the set of data created to support a specific short lived business situation?

(a)Personal Data Marts(b)Application Models(c)Downstream systems(d)Disposable Data Marts(e)Data mining models.

20.Which of the following does not form future access services?

(a)Authentication(b)Report linking(c)Push toward centralized services(d)Vendor consolidation(e)Web based customer access.

21.**What is the special kind of clustering that identifies events or transactions that occursimultaneously?**

(a)Affinity grouping(b)Classifying(c)Clustering(d)Estimating(e)Predicting.

22.Of the following team members, who do not form audience for Data warehousing?

(a)Data architects

(b)DBAs(c)Business Intelligence experts(d)Managers(e)Customers/users.

23.The precalculated summary values are called as

(a)Assertions(b)Triggers(c)Aggregates(d)Schemas(e)Indexes.

24.OLAP stands for

(a)Online Analytical Processing(b)Online Attribute Processing(c)Online Assertion Processing(d)Online Association Processing(e)Online Allocation Processing.

25.Which of the following employ data mining techniques to analyze the intent of a user query,provided additional generalized or associated information relevant to the query?

(a)Iceberg Query Method(b)Data Analyzer(c)Intelligent Query answering(d)DBA(e)Query Parser.

26.Of the following clustering algorithm what is the method which initially creates a hierarchicaldecomposition of the given set of data objects?

(a)Partitioning Method(b)Hierarchical Method(c)Density-based method(d)Grid-based Method(e)Model-based Method.

27.Which one of the following can be performed using the attribute-oriented induction in a mannersimilar to concept characterization?

(a)Analytical characterization(b)Concept Description.(c)OLAP based approach(d)Concept Comparison(e)Data Mining.

28.Which one of the following is an efficient association rule mining algorithm that explores the level-wise mining?

(a)FP-tree algorithm(b)Apriori Algorithm(c)Level-based Algorithm